Utilizing Antarctic Metagenomic Resources for the Identification of Hydrolases

Author:
Mohd Omar, Suhaila

Publication Date:
2014

DOI:
https://doi.org/10.26190/unsworks/16866

Utilizing Antarctic Metagenomic Resources for the Identification of Hydrolases

Suhaila Mohd Omar

A thesis in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Biotechnology and Biomolecular Science
Faculty of Science
University of New South Wales

May 2014

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Mohd Omar

First name: Suhaila

Other name/s:

Abbreviation for degree as given in the University calendar: PhD

School: School of Biotechnology and Biomolecular Science

Faculty: Faculty of Science

Title: Utilizing Antarctic metagenomic resources for the identification of hydrolases

Abstract 350 words maximum: (PLEASE TYPE)

Permanently cold environments are populated by a diversity of microorganisms that possess cold-adapted enzymes with potential biotechnological applications. Advances in molecular techniques have enabled bioprospecting of novel enzymes from extreme environments without tedious cultivation efforts. Cloning and sequencing of the metagenomic DNA from the water column of the meromictic Antarctic lake, Ace Lake, enabled enzyme identification by sequence-based searches and functional screening for enzyme activity. The sequence-based search for hydrolases using profile Hidden Markov Models indicated the presence of uncharacterized subtilases, lipases and glycoside hydrolases in the metagenomic dataset. Information from the bioinformatic analysis was utilized to perform targeted functional screening of clones containing the candidate hydrolase genes. Alongside this targeted approach, ~20000 clones from the Ace Lake metagenomic library were randomly screened via agar-based assay. As both functional screening approaches yielded very low positive identifications, the capacity of the sequence-based screening to detect *bona fide* hydrolases was further tested by cloning selected subtilase genes and overexpressing them in *Escherichia coli*. One of the overexpressed subtilase genes showed proteolytic activity when tested with N-succinyl-Ala-Ala-Pro-Phe-p-nitroanilide. Finally, in order to get a bigger picture of the abundance of subtilase and other peptidase genes in Ace Lake, taxonomic affiliation and relative abundance of all candidate peptidases were assessed. The Ace Lake peptidases composition was also compared to peptidases in Organic Lake and Southern Ocean metagenomic datasets. The analysis indicated an abundance of diverse metallopeptidases and serine peptidases in all three datasets. Peptidases related to protein and membrane regulatory control and chaperones with peptidase activity were the most abundant groups, which are linked to microbial survival as a response to environmental stress. In conclusion, this study revealed the advantage of cultivation-independent methods for bioprospecting enzymes from cold environments, and established a foundation for future functional studies of other enzymes.

Declaration relating to disposition of project thesis/dissertation

I hereby grant to the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all property rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstracts International (this is applicable to doctoral theses only).

Signature ..... Witness ..... 14/5/14 Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY

Date of completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed .....

Date 14/5/14

COPYRIGHT STATEMENT

'I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.
I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International (this is applicable to doctoral theses only).
I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.'

Signed .....

Date 2/7/14

AUTHENTICITY STATEMENT

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

Signed .....

Date 2/7/14

Abstract

Permanently cold environments are populated by a diversity of microorganisms that possess cold-adapted enzymes with potential biotechnological applications. Advances in molecular techniques have enabled bioprospecting of novel enzymes from extreme environments without tedious cultivation efforts. Cloning and sequencing of the metagenomic DNA from the water column of the meromictic Antarctic lake, Ace Lake, enabled enzyme identification by sequence-based searches and functional screening for enzyme activity. The sequence-based search for hydrolases using profile Hidden Markov Models indicated the presence of uncharacterized subtilases, lipases and glycoside hydrolases in the metagenomic dataset. Information from the bioinformatic analysis was utilized to perform targeted functional screening of clones containing the candidate hydrolase genes. Alongside this targeted approach, ~20000 clones from the Ace Lake metagenomic library were randomly screened via agar-based assay. As both functional screening approaches yielded very low positive identifications, the capacity of the sequence-based screening to detect *bona fide* hydrolases was further tested by cloning selected subtilase genes and overexpressing them in *Escherichia coli*. One of the overexpressed subtilase genes showed proteolytic activity when tested with N-succinyl-Ala-Ala-Pro-Phe-ρ-nitroanilide. Finally, in order to get a bigger picture of the abundance of subtilase and other peptidase genes in Ace Lake, taxonomic affiliation and relative abundance of all candidate peptidases were assessed. The Ace Lake peptidases composition was also compared to peptidases in Organic Lake and Southern Ocean metagenomic datasets. The analysis indicated an abundance of diverse metallopeptidases and serine peptidases in all three datasets. Peptidases related to protein and membrane regulatory control and chaperones with peptidase activity were the most abundant groups, which are linked to microbial survival as a response to environmental stress. In conclusion, this study revealed the advantage of cultivation-independent methods for bioprospecting enzymes from cold environments, and established a foundation for future functional studies of other enzymes.

Acknowledgement

Special thanks to my supervisor, Ricardo Cavicchioli and co-supervisor, Timothy Charlton for the advice, support and guidance throughout my candidature. I am very grateful for the time, energy, patience and hard work that were spent for me.

To Tim William, Jodi Richards, Sheree Yau and David Wilkins, all of you are wonderful friends, teachers, readers and motivators for me. Tim, thanks for being patient for all the lab and editorial requests. Jodi, it has been a short time, but I believed God has sent you to let me keep going. Thank you for everything. Sheree, you are the most patient teacher that I have ever met. Thanks for reading my thesis and all the UNIX tips. David, when working in the lab on quiet weekend, it was good to know there was someone else on the same floor, and thanks for your tips for data normalization. That were just a few, and I believed, all of you deserve my appreciation for many other reasons.

Sohail Siddiqui, we might disagree on many things but you have taught me well about enzymes. It is funny though to share similar name with a man in the same lab. Federico Lauro, Matt DeMaere and Torsten Thomas, thank you for all the help related to bioinformatics. For the other Cavicchioli lab members, Haluk Ertan, Davide De Francisci, Kevin Chong, Michelle, Tahria, Taha, Arjun, it has been a memorable years in 313 lab.

Special thanks to Yuslina Zakaria, it is always good to have a friend that understand computer language. Thanks for your help with the scripts and for being very understanding listener. I appreciate our friendship very much.

Thanks to Anne Poljak from Bioanalytical Mass Spectrometry Facility (BMSF) UNSW, for help with gel processing and mass spectrometry analysis.

Melanie, I am fortunate to have you as friends next door during this PhD journey, to share stories, from experimental works to family issues.

Thank you to Ministry of Higher Education of Malaysia and International Islamic University of Malaysia for financially supported my study abroad. Thanks to my IIUM-family, especially to Sis. Zarina Zainuddin, Zaima Azira, Anil Azura and Nuraslinda for the support when I was writing my thesis.

To Mak, Bapak, Sarah Syukri and Safiah, thank you for always being there for me. I am grateful to be the part of this family.

To my other half Khairul Azmi, thank you for all your patience, love and support. To all my children; Ahmad Syahid, Alya Syifa and Ammar Syafi-you all are the source of my strength throughout this journey. Thanks for making every day a day to look forward to.

Table of Contents

ABSTRACT ..... **IV**

ACKNOWLEDGEMENT ..... **V**

TABLE OF CONTENTS ..... **VII**

LIST OF FIGURES ..... **XIII**

LIST OF TABLES ..... **XVI**

ABBREVIATIONS ..... **XVIII**

CHAPTER 1: GENERAL INTRODUCTION **1**

1.1 Life at low temperatures 1

1.2 Microbial diversity in low temperature environments 2

1.3 Biotechnological potential of cold-adapted enzymes 4

1.4 Unravelling novel cold-adapted enzymes via metagenomics 6

1.4.1 Function-based screening 7

1.4.2 Challenges and improvements of functional screening 13

1.5 Sequence-based screening 14

1.6 Ace Lake: Potential resources of cold-adapted enzymes 17

1.7 Objectives 19

CHAPTER 2: SEQUENCE-BASED SCREENING OF HYDROLASES IN THE ACE LAKE METAGENOMIC DATASETS USING PROFILE HIDDEN MARKOV MODELS **22**

2.1 Introduction 22

2.1.1 Metagenomics and microbial life in cold environments 22

2.1.2 Metagenomics of Ace Lake 23

2.1.3 Sequence-based screening of the metagenomic data using profile Hidden Markov Models 25

2.1.4 Classification of hydrolases based on sequence homology 25

2.2 Materials and methods 27

2.2.1 Ace Lake samples 27
2.2.2 Dataset description 28
2.2.3 HMMsearch for hydrolases from Ace Lake metagenome dataset....29
2.2.4 Selection of the sequences from the HMMsearch results....29
2.2.5 Annotation of putative hydrolase sequences and domain architecture analysis....30
2.2.6 Phylogenetic analysis....30
2.3 Results....30
2.3.1 Distribution of the HMMsearch result for subtilase, lipases and glycoside hydrolases 30
2.3.2 Selection of the subtilase sequences from the HMMsearch results 34
2.3.3 Lipases sequences in the HMMsearch results 40
2.3.4 Selection of the GH13 sequences in the HMMsearch results 46
2.4 Discussion....54
2.4.1 Implication of DNA sequencing and processing techniques towards the number of matches in the HMMsearch 54
2.4.2 The matches of hydrolases were linked to the dominant taxa in the Ace Lake environment 55
2.4.3 Conserved domain architecture of subtilase inferred probable biological function of the enzymes 56
2.4.4 Functional potential of lipase GDSL in the Ace Lake aquatic environment 59
2.4.5 Trehalose synthase and its role for viability at low temperatures 59
2.5 Conclusion 61

CHAPTER 3: FUNCTIONAL SCREENING OF HYDROLASES FROM THE ACE LAKE METAGENOMIC CLONE LIBRARY 62

3.1 Introduction 62
3.2 Materials and methods 63
3.2.1 Library construction 63

3.2.2 Identification of the targeted gene from the sequence-based screening in the source clone library....64

3.2.3 Agar-based functional screening of hydrolases from the Ace Lake metagenomic clone library....64

3.2.4 Enzymatic assay....65

3.2.5 Zymography....67

3.2.6 Sequence analysis....68

3.3 Results....68

3.3.1 Agar-based assay....68

3.3.2 Liquid assay....71

3.3.3 Zymography....75

3.3.4 Sequence analysis....76

3.4 Discussion....78

3.4.1 Functional screening of the metagenomic library....78

3.4.2 Low hit rates of hydrolase activity in the Ace Lake metagenomic library....79

3.4.3 The pitfalls of using skimmed milk and tributyrin agar-based assay to detect proteolytic and lipolytic activity in the metagenomic clone library....80

3.5 Conclusion....82

CHAPTER 4: CLONING AND EXPRESSION STUDIES OF THE SELECTED SUBTILASE GENES FROM THE ACE LAKE METAGENOMIC SEQUENCING LIBRARY....84

4.1 Introduction....84

4.2 Materials and methods....86

4.2.1 Physico-chemical analysis....86

4.2.2 Structural modelling....86

4.2.3 Cloning of subtilase genes....87

4.2.4 Competent cell preparation....88

4.2.5 Cell transformation....89

4.2.6 Identification of positive clones....89

4.2.7 DNA sequencing 90
4.2.8 Heterologous expression of recombinant proteins 91
4.2.9 Protein sample preparation 91
4.2.10 Determination of protein concentration 92
4.2.11 SDS-polyacrylamide gel electrophoresis (SDS-PAGE) 92
4.2.12 Solubilization of protein that was expressed as inclusion bodies 93
4.2.13 Protein identification by mass spectrometry 96
4.2.14 Enzymatic assays 96
4.3 Results 97
4.3.1 Analysis of 163539195 (Subt9195) 97
4.3.2 Analysis of 163128715 (Subt8715) 112
4.3.3 Analysis of 167865372 (Subt5372) 116
4.4 Discussion 122
4.4.1 Protein properties and overexpression in pET expression system 122
4.4.2 Extracellular localization and autoprocessing of protease propeptide region....123
4.4.3 Solubilization of Subt9195 protein 124
4.4.4 Unsuccessful subtilase maturation mechanism effects towards protease activity 126
4.4.5 Effect of temperature towards Subt8715 protease activity 128
4.4.6 Prediction of tertiary structure of Subt9195 and Subt8715 129
4.5 Conclusion 130

CHAPTER 5: COMPARATIVE PEPTIDASE ANALYSIS OF ACE LAKE, ORGANIC LAKE AND SOUTHERN OCEAN METAGENOMIC DATA 131

5.1 Introduction 131
5.1.1 Antarctica: Potential resources of cold-adapted enzymes 132
5.2 Materials and methods 134
5.2.1 Data description 134

5.2.2 Calculating peptidase associated reads abundance in Ace Lake, Organic Lake and Southern Ocean metagenome dataset 135

5.2.3 Data normalization 136

5.2.4 Data analysis 136

5.3 Results 137

5.3.1 Analysis of COG annotations associated to peptidases in the Ace Lake, Organic Lake and Southern Ocean metagenome 137

5.3.2 Variation of taxa associated with peptidase genes in Ace Lake, Organic Lake and Southern Ocean 143

5.3.3 The abundance of metallopeptidase and contribution of the dominant phyla in the sample 145

5.3.4 Comparison of COG1404 abundance in the samples 149

5.4 Discussion 150

5.4.1 Dominant phyla contribution to the abundance of metallopeptidase 150

5.4.2 The prevalence of chaperone, protein regulatory control and membrane biogenesis peptidase in the metagenomic samples 151

5.4.3 Comparison of subtilisin abundance in the samples 152

5.4.4 Abundance of oligo, di and tripeptidyl peptidase 153

5.4.5 Abundance of peptidase E and imidazolonepropionase in the Organic Lake and Southern Ocean 154

5.5 Conclusion 154

CHAPTER 6: FUTURE PERSPECTIVES AND CONCLUSIONS 155

6.1 Introduction 155

6.2 Possible future work for bioprospecting of enzymes from Antarctic metagenome 155

6.2.1 High-throughput expression and purification systems.....156

6.2.2 The advantage of multiple-host metagenomics expression system 157

6.2.3 The potential success of heterologous expression using synthetic DNA 158

6.3 Concluding remarks 161

REFERENCES **162**

APPENDIX A **183**

APPENDIX B **186**

APPENDIX C **188**

APPENDIX D **196**

APPENDIX E **197**

APPENDIX F **200**

List of Figures

Figure 1.1: Overview of the metagenomic screening process..... 7

Figure 1.2: A map of the Vestfold Hills showing fjords, bays and lakes..... 18

Figure 2.1: Physicochemical and biological structure of the Ace Lake..... 24

Figure 2.2: Distribution of subtilase (left) and amylase (right) in percentage number of counts relative to the number of translated ORFs in the metagenome datasets…….32

Figure 2.3: Distribution of lipase GDSL (left) and lipase 3 (right) in percentage number of counts relative to the number of translated ORFs in the metagenome datasets…….33

Figure 2.4: Distribution of the subtilase sequence matches in the 0.1 µm data……35

Figure 2.5: The phylogenetic tree of the subtilase-like protein sequences from Ace Lake as inferred using the Neighbor-Joining method.....38

Figure 2.6: Distribution of lipase GDSL sequences matches in the 0.1 µm data.....42

Figure 2.7: Phylogenetic tree of the Ace Lake metagenome-derived lipase GDSL sequences.....44

Figure 2.8: Distribution of GH13 sequence matches in the 0.1 µm data.....48

Figure 2.9: Phylogenetic three of the GH13 sequences derived from the Ace Lake metagenome.....52

Figure 3.1: Flow chart of the enzyme screening process..... 63

Figure 3.2**:** Results for agar-based screening for protease activity……69

Figure 3.3**:** Results for agar-based screening for amylase activity on the starch agar.....70

Figure 3.4: Trybutyrin agar assay after 2 weeks incubation at 25°C.....71

Figure 3.5: Confirmation data of the protease activity of clone D24 in azocasein assay…..72

Figure 3.6**:** Growth curves (OD 650 nm) for the selected positive clones (D38 & D8) for amylase activity and the negative control (T).....73

Figure 3.7**:** Enzyme activity assay (α-amylase) for intracellular extracts of the selected clones compared to the host and negative control using DNS assay.....74

Figure 3.8: Result of the filter paper 4-MUF-butyrate assay……74

Figure 3.9: Results of Native PAGE zymogram for lipase activity ……75

Figure 3.10:Overview of the read assembled in the scaffold 7180000121853.....76

Figure 3.11: Schematic overview of the annotation in the scaffold 7180000121853.....76

Figure 3.12**:** Graphical overview of the reads assembly of part of the scaffold 7180000134253..... 77

Figure 3.13**:** Schematic overview of the annotation in the scaffold 7180000134253 at locus 27271 towards 33363..... 77

Figure 4.1 : Procedures for the attempt to solubilize the protein that was expressed as IBs....93

Figure 4.2: Kyte and Dolittle hydropathy plot of Subt9195 protein....98

Figure 4.3: SDS-PAGE of total cell protein in expression studies of Subt9195 at 25°C in *E. coli* BL21 (DE3)....99

Figure 4.4: SDS-PAGE of the concentrated extracellular protein fraction from expression studies of Subt9195 at 25°C in *E. coli* BL21 (DE3)....99

Figure 4.5: SDS-PAGE of the soluble fraction from expression studies of Subt9195 at 25°C in *E. coli* BL21 (DE3)....100

Figure 4.6: SDS-PAGE of total cell protein from expression studies of Subt9195 at 30°C in *E. coli* BL21 (DE3)....100

Figure 4.7: Growth profiles of pET28bSubt9195 in the expression studies at 10°C ……

Figure 4.8: SDS-PAGE of expression studies of Subt9195 at 10°C....102

Figure 4.9: Skimmed milk agar assay of culture Subt9195 at 25 and 30°C....103

Figure 4.10: SDS-PAGE of solubilization of Subt9195 protein in 8M urea....104

Figure 4.11: SDS-PAGE of solubilization of IB of Subt9195 following prewashed with Triton X-100, 1M urea and 0.05% sarkosyl. 105

Figure 4.12: Peptides fragments that matched to the deduced amino acid sequence of Subt9195 in the FTMS analysis....106

Figure 4.13: SDS-PAGE of soluble fraction of pET43aSubt9195NusHisC in the expression studies at 30°C....107

Figure 4.14: SDS-PAGE of soluble fraction of pET43a.1 culture inducted with 1mM IPTG at 30°C 108

Figure 4.15: SDS-PAGE of fractions obtained after pET43a_Subt9195NusAHisC His-tag purification....109

Figure 4.16: Purified fractions of pET43a_Subt9195NusAHisC before and after concentration and dialysis; on 4-12% Bis–Tris NuPage (Invitrogen) gel.. 110

Figure 4.17: SDS-PAGE of fraction 4 after 2 weeks storage at 4°C indicate self-digesting activities....110

Figure 4.18: Protease activity of His-tag purified Subt9195NusAHisC at room temperature. 111

Figure 4.19: Kyte and Dolittle hydropathy plot of Subt8715....112

Figure 4.20: SDS-PAGE of soluble fraction of pET28bSubt8715 in BL21 (DE3) at 25 and 30°C. 113

Figure 4.21: SDS-PAGE of concentrated extracellular protein fraction from expression studies of pET28bSubt8715 in *E. coli* BL21 (DE3) 113

Figure 4.22: Activity assay of crude cell extracts of Subt8715 compared to the negative control (pET28b)....114

Figure 4.23: Effect of the temperature on Subt8715 activities towards AAPF.....115
Figure 4.24: Ln OD vs 1/T of Subt8715 115
Figure 4.25: Kyte and Dolittle hydropathy plot of Subt8715 116
Figure 4.26: SDS-PAGE of total cell protein and soluble fraction of pet28bSubt5372 in Rosetta 2 (DE3) at 25 and 30°C. 117
Figure 4.27: Predicted 3D model of Subt9195 from different angle. 118
Figure 4.28: Structural alignment of Subt9195 and 3AFGA.....119
Figure 4.29: Predicted 3D model of Subt8715 120
Figure 4.30: Structural alignment of Model 1 and Model 2 with 1R6VA.....121
Figure 5.1: Hierarchical clustering based on COG category abundances. 141
Figure 5.2: Statistical (STAMP) analysis of normalized counts of COG annotated proteins among Ace Lake, Organic Lake and Southern Ocean. 142
Figure 5.3: Relative abundance of taxa associated with peptidase genes in Ace Lake, Organic Lake and Southern Ocean samples at phyla level.....143
Figure 5.4: Hierarchical clustering based on taxonomic associations of KEGGs hits to the peptidase genes. 144
Figure 5.5: Comparison of peptidase types distribution in: a) Combination of Ace Lake, Organic Lake and Southern Ocean data and b) Global Ocean Sampling expedition samples (GOS60). 145
Figure 6.1: Productivity in DNA synthesis and sequencing using commercially available instruments……159
Figure 6.2 : The procedure developed at DNA 2.0 for designing a gene sequence to encode a specific protein.....160
Figure E.1: Multiple alignment of the deduced amino acid sequences of Subt9195 and Subt5372 with matches from NCBI and Swiss-Prot database.....197
Figure E.2: Multiple sequence alignments of the deduced amino acid sequences of Subt8715 (766aa) and matched sequences from NCBI and Swiss-Prot databases...198
Figure E.3: Multiple alignment of the deduced amino acid sequences of Subt4518 and with matches from NCBI and Swiss-Prot database.....199

List of Tables

Table 1.1: Metagenome-derived cold-adapted enzymes isolated via functional and sequence-based screening 8
Table 1.2: List of new generation sequencing platforms and their expected throughputs, error types and error rates 15
Table 2.1: Summary of the Ace Lake metagenomic datasets used in the sequence-based screening 28
Table 2.2: List of the selected profile HMM utilized in the sequence-based screening process 29
Table 2.3: The breakdown of the types of hydrolases identified in the Ace Lake metagenomic dataset with default E-value cut-off of 10 31
Table 2.4: List of the subtilase matches that passed the high stringency filtering criteria based on E-value, protein length, presence of catalytic sites and complete ORF prediction 36
Table 2.5: Result of analysis of domain architecture for selected subtilase sequences 39
Table 2.6: List of the lipase 3 sequences that passed the high stringency filtering criteria based on E-value, protein length, presence of catalytic sites and complete ORF prediction 40
Table 2.7: List of the lipase GDSL sequences that passed the high stringency filtering criteria based on E-value, protein length, presence of catalytic sites and complete ORF prediction 41
Table 2.8: Result of analysis of domain architecture for selected lipase GDSL sequences. 45
Table 2.9: List of the GH13 sequences that passed the higher stringency filtering criteria based on E-value, protein length, presence of catalytic sites and complete ORF prediction. 47
Table 2.10: Analysis of domain architecture of the GH13 sequences 53
Table 3.1: The result for agar-based screening of a metagenomic DNA library for different enzymatic activities 68
Table 4.1: Primers used in PCR amplification of selected subtilase genes. Restriction sites are shown in italics and underlined. 88
Table 4.2: Primers used for sequencing 90
Table 4.3: List of denaturants, solubilising agents and refolding buffers used in the protein solubilization and refolding process 94
Table 4.4: Estimated accuracy of the structure model 118
Table 4.5: Estimated accuracy of the structure model 119
Table 5.1: Classification of peptidase 132
Table 5.2: List of the datasets used in the metagenomic analysis. 135
Table 5.3: COG ID and count of COG annotation related to peptidase from Ace Lake, Organic Lake and Southern Ocean 138
Table 5.4: Contribution of different taxonomic group to counts of COG categories related to metallopeptidase in Ace Lake 146
Table 5.5: Contribution of different taxonomic group to counts of COG categories

related to metallopeptidase in Organic Lake……147
Table 5.6: Contribution of different taxonomic group to counts of COG categories related to metallopeptidase in Southern Ocean……148
Table 5.7 : Contribution of different taxonomic group to counts of COG1404 (Subtilisin)……149
Table C.1 Annotation of the HMMsearch result for subtilase……188
Table C.2: Annotation of HMMsearch result for Lipase 3……190
Table C.3: Annotation of the HMMsearch result for lipase GDSL……191
Table C.4: Annotation of the HMMsearch results for GH13……193
Table D.1 :Final concentration of ammonium sulfate……196
Table F.1: The spreadsheet for identification of the targeted clone from the sequence-based screening in the metagenomic clone library……200

Abbreviations

α	alfa
β	beta
γ	Gamma
&	and
µ	Micro
%	percentage
°C	degree Celsius
µm	micrometer
aa	amino acid
AAPF	succinyl-AAPF-ρ-nitroanilide
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
bp	base pair
BSA	Bovine serum albumin
$CaCl_2$	Calcium chloride
CDart	Conserved Domain Architecture Retrieval Tool
CDD	Conserved Domain Database
CH_4.	Methane
CHAPS	[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate
cm	centimeter
CO_2	Carbon dioxide
ddH_2O	distilled water
DNA	deoxyribonucleic acid
DNS	dinitrosalicylic Acid
DOC	dissolved organic carbon

DTT	Dithiothreitol
E_a	activation energy
EDTA	Ethylenediaminetetraacetic acid
g	Gram/gravity
GFP	Green fluorescent protein
GH13	Glycoside hydrolase family 13
GRAVY	Grand average of hydropathicity index
GST	glutathione S-transferase
h	hour
HEPES	4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid
His	histidine
HMM	Hidden Markov Model
HSL	hormone sensitive lipase
i.e	id est (“that is”)
IB	inclusion body
IMAC	immobilised metal ion chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kb	kilo base pairs
kDa	kilo Dalton
KEGG	Kyoto Encylopedia of Genes and Genomes
kg	kilogram
kJ/mol	kilo Joule per mole
LB	Luria Bertani
m	meter
M	molar
m^2	meter square
mg	miligram
$MgCl_2$	Magnesium chloride

min	minute
mL	mililiter
mm	milimeter
MPB	maltose-binding protein
mRNA	messenger ribonucleic acid
MUSCLE	**MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation
n.a	not available
NCBI	The National Center for Biotechnology Information
nr	non-redundant
NEB	New England Biolabs
nm	nanometer
NGS	Next Generation Sequecncing
NusA	Transcription elongation protein NusA
OM	Outer membrane
ORF	Open reading frame
Pal	Peptidoglycan-associated lipoprotein
PIGEX	Product-induced gene expression
PMSF	Phenylmethylsulfonyl fluoride
PorSS	Por secretion system
PPC	Peptidase C-terminal domain
rpm	revolution per minute
rRNA	Ribosomal ribonucleic acid
Sarkosyl	sodium lauroyl sarcosinate
SDS	sodium dodecyl sulfate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SIGEX	Substrate-induced gene expression screening
TreS	Trehalose synthase

CHAPTER 1
General Introduction

1.1 Life at low temperatures

Extremophiles, organisms that can thrive in the harshest environment on earth, are widely distributed in nature. Among them, thermophiles and hyperthermophiles are capable of growing at very high temperatures, and are found in places like hot springs, volcanic vents under water, and even in industrial chimneys that exhaust hot air. Psychrophiles, on the other hand, grow at very low temperatures. Psychrophiles, a term derived from the Greek words 'psychros', meaning cold, and 'philos' meaning loving, i.e., cold-loving, have successfully colonized all permanently cold environments from the deep sea to mountain and polar regions (D'Amico *et al.,* 2006). In fact, they are even found in domestic refrigerator and super market freezers, where they are well known for spoiling foods and drinks (Sørhaug and Stepaniak, 1997, Ercolini *et al.,* 2009). The term, 'psychrophiles' was first proposed by Schmidt-Nielson in 1902 to describe microorganisms that had ability to grow at 0°C (Ingraham and Stokes, 1959, Russell, 2003, Hoyoux *et al.,* 2004). Another related term commonly used in the food industry, 'psychrotroph', means 'cold-eating', was introduced by Eddy in 1960, to describe microorganisms that can spoil or poison refrigerated food. The term, 'psychrotolerant' is used in ecology to describe the thermal ability of this group (Russell, 2003).

The terminology to describe microorganisms able to grow at low temperatures has been debated and has changed over time. Before the 1960s, there was little consensus concerning the maximum, minimum, and optimum growth temperatures for psychrophilic microorganisms and much of the literature dealing with psychrophiles refer to microorganisms that are cold-tolerant rather than truly "cold-loving" (Morita, 1975, Zhang, 2008). The additional source of difficulty for defining psychrophiles was introduced by dairy bacteriologists for whom psychrophilic bacteria are those which can grow relatively rapidly in the range of 1.7 to 10°C, even though the upper part of this range includes mesophilic bacteria (Ingraham and Stokes, 1959). Ingraham and Stokes, introduced a definition that was not based on cardinal temperatures, stating that psychrophiles are microorganisms which grow rapidly enough at 0°C to form macroscopically visible colonies in about one or two weeks, which equates to a generation time at 0°C of less than 48 hours (1959). Later, Morita suggested a more

specific definition for psychrophiles as organisms having an optimum growth temperature of about 15°C or lower, a maximum temperature for growth at 20°C or below (1975). The term psychrotrophic, previously referred to as facultatively psychrophilic, is used for cold-tolerant organisms with maximum growth temperature above 20°C (Morita, 1975). Other proposed terms based on cardinal growth temperatures are 'eurypsycrophile' or 'eurythermal 'and 'stenopsychrophile' or 'stenothermal' (Feller and Gerday, 2003, Cavicchioli, 2006). The term stenopsychrophile (formerly 'true psychrophile') describes a microorganism with a restricted growth temperature range that cannot tolerate high temperatures for growth. Eurypsychrophile (formerly 'psychrotolerant' or 'psychrotroph' describes a microorganism that likes permanently cold environments, but can tolerate a wide range of temperatures extending into the mesophilic range (Cavicchioli, 2006).

Nevertheless, because cardinal growth temperatures cannot fully capture the adaptability of a microorganism to its environment and are unavailable for the vast majority of microbes in nature that evade cultivation, the term psychrophilic remains most useful for categorizing cultured microorganisms. Cold-adapted is used very generally to describe microorganisms, cultured or uncultured, that express a recognisable adaptation to low temperature, while cold-active is often reserved for enzymes and viruses with catalytic or infective activity at low temperature (Deming, 2009). In this thesis, 'cold-adapted' term is being used to describe both psychrophilic microorganisms and enzymes that active at low temperature.

1.2 Microbial diversity in low temperature environments

The earliest known report of microbial life in a cold environment dates back to the fourth century BC and the writings of the Greek philosopher Aristotle who made observations of the red algae that turned snow to reddish colour (Deming, 2009). More than two millennia later (1840), a botanist, Hooker, observed algal growth associated with coloured sea ice (Russell, 2003). This was followed by Certes in 1884, who reported the existence of bacteria growing in sediments at low temperatures (Russell, 2003). In 1887, the German scientist Forster was the first to describe the ability of a bioluminescent bacterium, derived from fish preserved in the cold, to reproduce at 0°C (Deming, 2009).

Identification of psychrophiles and understanding the mechanisms of cold adaptation in the past was hampered by limited technologies and facilities to obtain and preserve samples. Extensive studies of psychrophiles and their habitats in the past three

decades (Hoover and Pikuta, 2010), has contributed to the understanding that cold adaptation is not exclusively about temperature but related to other parameters including pressure, salinity, and nutrient availability (Deming, 2009). Viable bacteria, mainly dominated by Gram-positive, have been found in the atmosphere as high as the stratosphere and mesosphere (41-77km) where the temperature may reach as low as-100°C (Wainwright *et al.,* 2003, Griffin, 2008).

For a long time, glacier ice was considered to be devoid of life due to the harsh environment i.e; subfreezing temperatures ranging from -10 to -56°C, high hydrostatic pressure, and low nutrient, water and light availability (Price, 2007). However, this view changed with increasing studies on the microbial ecology and diversity of natural icy habitats, such as permanently ice-covered lakes (Priscu *et al.,* 1998, Takacs *et al.,* 2001), subglacial environments (Lanoil *et al.,* 2009) and glacier ice (Price, 2007, Simon *et al.,* 2009b). The total number of bacterial cells in the Antarctic and Greenland ice sheets was measured to be $9.16x10^{25}$, which corresponds to a significant carbon pool of $2.65x10^{-3}$Pg and represents a considerable reservoir of microbial diversity (Priscu and Christener, 2004, Priscu *et al.,* 2007).

Gilichinsky *et al.* (2005) have shown evidence of an unsuspected biodiversity in the sodium-chloride water brines (cryopegs) derived from ancient marine sediments and sandwiched within permafrost. The water brines have remained liquid for about 100,000 years at -9 to -11°C. Sequence analysis of the 16S rRNA genes showed that the 17 phylotypes were most closely related (>95% sequence similarity) to the following nine genera: *Psychrobacter*, *Arthrobacter*, *Frigoribacterium*, *Subtercola*, *Microbacterium*, *Rhodococcus*, *Erwinia*, *Paenibacillus*, and *Bacillus*. Currently, the lowest recorded growth temperature is of Arctic sea ice bacterium, *Psychromonas ingrahamii* (Moyer and Morita, 2001, Breezee *et al.,* 2004, Margesin *et al.,* 2007).

Among the psychrophilic bacteria that have been detected, the most commonly reported are the Gram-negative *α*-, *β*-and *γ-Proteobacteria*, and members of the *Bacteroidetes* (formerly known as *Cytophaga–Flavobacterium–Bacteriodes* phylum) while *Coryneforms*, *Arthrobacter* and *Micrococcus* are the most frequently found Gram-positive bacteria (D'Amico *et al.,* 2006). The genera of psychrophilic archaea that are often reported are *Methanogenium*, *Methanococcoides* and *Halobium* (Feller and Gerday, 2003). Bacteria generally dominate in number and diversity over archaea, although in some areas such as deep sea waters, these are found in equivalent numbers, with

Methanogenium and *Methanococcus* being the most cited genera (D'Amico *et al.,* 2006). Among identified *Cyanobacteria*; *Oscillatoria*, *Phormidium* and *Nostoc* were dominant in most of the investigated Antarctic habitats (Pandey *et al.,* 2004). Psychrophilic yeasts, such as *Candida* spp., *Leucosporidium* spp. and particularly *Cryptococcus* spp., have been isolated repeatedly from various sites in Antarctica (Kingston and Brent, 2001). The increased application of molecular techniques based on 16s rRNA genes to polar and cold deep sea environments has created substantial databases on the diversity of bacteria and archaea inhabiting cold waters, sediments and ice (Kimhi and Magasanik, 1970, Brown and Bowman, 2001, Smith and Magasanik, 1971, Lauro *et al.,* 2011).

1.3 Biotechnological potential of cold-adapted enzymes

Over the last few decades, studies on cold-adapted microbes have increased considerably, which can be attributed to several factors, such as the awareness of accelerated environmental changes in polar regions, a strong interest in the habitability of frozen areas elsewhere in the universe (astrobiology), and a realization of the considerable biotechnological potential of these organisms (Cavicchioli *et al.,* 2002, Margesin *et al.,* 2002, de Pascale *et al.,* 2012). A number of companies are mining the biological resources of polar regions for their biotechnological potential and the number of patents in this field is growing (Hannah, 2008).

Low temperatures and freezing conditions influence the lives of all organisms in multiple ways. For example low temperatures increase viscosity of the medium, change membrane fluidity and protein conformation, nutrient availability, ability to successfully reproduce, and introduces the need for protection against freezing (Margesin *et al.,* 2007). Perhaps the most significant challenge faced by psychrophiles is to maintain appropriate rates of enzyme-catalysed reactions involved in essential cellular process at low temperatures. Psychrophiles address this challenge by synthesizing enzymes with high catalytic efficiency (K_{cat} /K_m) at low and moderate temperatures (0–30°C) at which homologous enzymes produced by microorganisms from other thermal classes have little or no activity (Margesin *et al.,* 2007). This characteristic, if harnessed for biotechnological use, offers potential economic benefits, for example, through substantial energy savings in large scale processes that would otherwise require expensive heating reactors. A typical example is the 'peeling' of leather by proteases which can be done at tap water temperature by cold-active enzymes instead of heating to 37°C, for the process to be performed by mesophilic enzymes (Feller and Gerday, 2003). The use of cold-adapted hydrolytic enzymes such as protease, lipase, amylase and cellulase in the formulation of

detergent is a great advantage for cold washing as it not only eliminates the need for heating but also reduce the wear and tear of textile fibers (Gomes and Steiner, 2004).

Furthermore, cold-adapted enzymes are heat labile and are frequently inactivated at temperatures that are not detrimental for their mesophilic counterparts and therefore they can be efficiently inactivated by moderate heat input (Margesin and Feller, 2010). The use of heat labile alkaline phosphatase in molecular biology, which was commercialized by New England Biolabs (Cavicchioli *et al.,* 2011), is an excellent alternative to *Escherichia coli (E. coli)* and intestinal alkaline phosphatase. Since the heat labile enzyme is inactivated by moderate heat treatment, the subsequent steps can be done in the same tube and minimized nucleic acid losses (Margesin and Feller, 2010). Apart from these uses, cold-adapted enzymes are known for their use in the food industry. Two examples are, the removal of lactose from milk by psychrophilic β-galactosidase during cold storage, a process which has been patented (Hoyoux *et al.,* 2001), and the use of cold-active pectinases to reduce viscosity and clarify fruit juice at low temperatures (Feller and Gerday, 2003). In the baking industry, the use of psychrophilic glycosidase improves the quality of the final products as mesophilic glycosidase used in the baking process can retain residual activity after cooking that alters the structure of the final product during storage (Collins *et al.,* 2006).

More novel enzymes are expected to be found from genetic resources in cold environments. For example a novel cold-adapted cellulase from an earthworm living in a cold environment was discovered with the ability to convert cellulose directly into glucose (Ueda *et al.,* 2010). This might be an important discovery for efficient production of biofuels from cellulosic waste at low temperatures (de Pascale *et al.,* 2012). In a survey of patents related to Antarctica, lipases from *Candida antarctica* by far dominate the number of process or product-based patents (Lohan and Johnston, 2005), which suggests the potential of enzymes from cold-adapted microorganisms.

Isolating microorganisms by employing selective media and cultivation conditions for phylogenetic groups known for producing interesting bioactivities is still the basis for most microbiological bioprospecting efforts in cold-adapted environments. More sophisticated media and cultivation conditions (Vartoukian *et al.,* 2010), as well as physical access to new habitats (e.g. from glacial and permafrost core samples or underneath sea ice by diving) (Loveland-Curtze *et al.,* 2010), have steadily brought new phylotypes of cold-adapted microorganisms into the stock culture collections. The

application of molecular, culture-independent methods for assessing microbial diversity has shown that the diversity of the uncultured majority is huge (Torsvik *et al.,* 1990, DeLong and Pace, 2001) and offer a great deal of biocatalytic potential. However, most microorganisms in the environment are not readily cultivated (Pace *et al.,* 1986, Amann *et al.,* 1995, Hugenholtz *et al.,* 1998), which limits investigation into this large genetic resource.

1.4 Unravelling novel cold-adapted enzymes via metagenomics

An extension to the possibilities offered by cultivation has come with the emergence of metagenomics. Metagenomics is defined as the study of the pooled genetic complement of a given environmental sample, where analysis can be either sequence driven or function driven (Handelsman, 2004). The field of metagenomics developed from advances made in DNA extraction and cloning from environmental samples (Schmidt *et al.,* 1991). The construction of metagenomic libraries and other DNA-based metagenome projects are initiated by isolation of high quality DNA that is suitable for cloning and covers the microbial diversity present in the original habitat (Simon and Daniel, 2011). The advances in sequencing technologies and developments in high-throughput molecular biology techniques, such as robotized handling of large numbers of clones and screening of bioactivity, have paved the way for metagenomics approaches to microbial bioprospecting.

Most researches use *E. coli* as a surrogate host for metagenomic library construction and various types of *E. coli* strains are available from commercial sources. The choice of vector depends largely on the length of insert. Plasmids are suitable for cloning smaller than 10kb DNA fragments; cosmids (23-35kb), fosmids (25-40kb) or BACs (100-200kb) can be used to clone larger fragments (Uchiyama and Miyazaki, 2009). Metagenomic libraries of DNA extracted directly from environmental samples provide genomic sequences with phylogenetic and functional information. In principle, as illustrated in Figure 1.1, the techniques for recovery of novel biomolecules using metagenomics can be divided into two main approaches: function-based and sequence-based screening of metagenomic libraries.

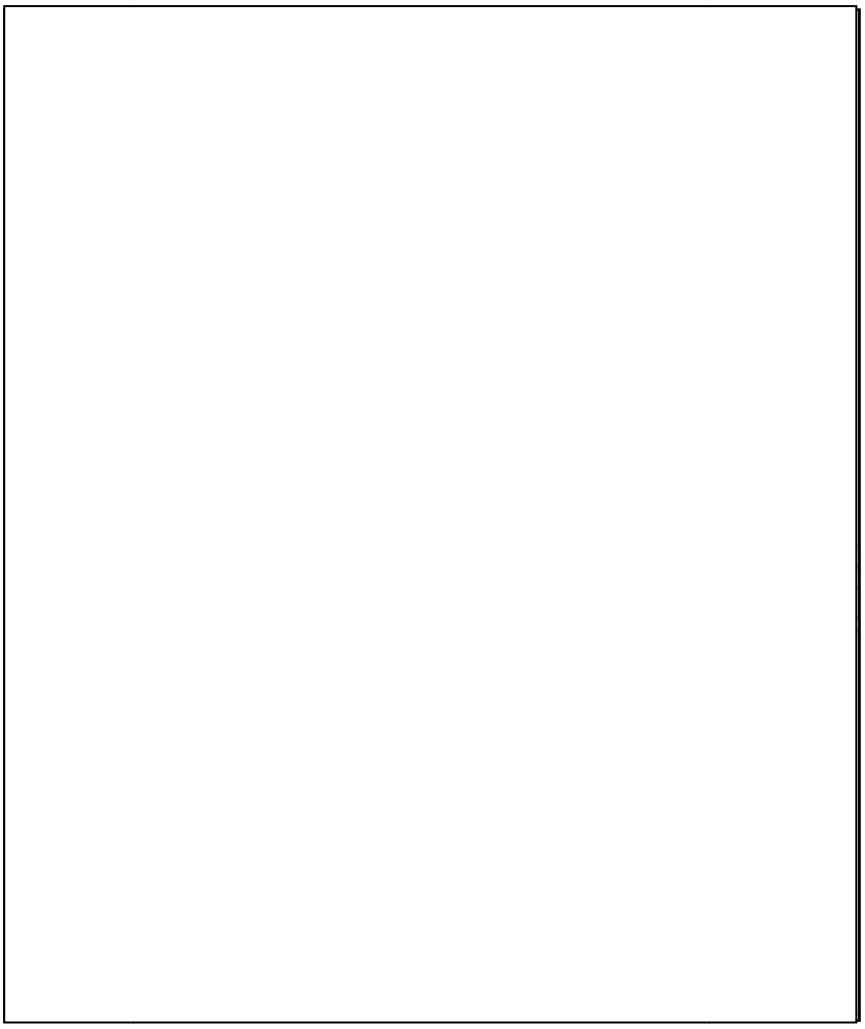


Figure 1.1: Overview of the metagenomic screening process (Iqbal *et al.,* 2012). Figure has been removed due to Copyright restrictions.

1.4.1 Function-based screening

Most of the screening methods for the detection of genes encoding novel biomolecules are based on the metabolic activities of the metagenomic-library-containing clones. The power of this approach is that it does not require the genes of interest to be recognisable by sequence analysis and thus has potential to directly identify entirely new classes of genes for both known and novel functions (Handelsman, 2004). Recent studies on function-based screening of metagenomic libraries that resulted in cold-adapted enzymes are summarised in Table 1.1.

Table 1.1: Metagenome-derived cold-adapted enzymes isolated via functional and sequence-based screening (adapted from Cavicchioli *et al.,* 2011 with modifications).

Enzyme	Environment	Host/ Vector	Positive Clones/ Total Clones	Screening Technique/ Substrate	T_{opt} (°C)	pH_{opt}	Level Of Characterization	Reference
Lipase	Baltic sea sediment	*E. coli*/ fosmid	70/ >7000	Agar-based assay –1% Tributyrin and 0.1% gum arabic	35	na	Protein purification, temperature, substrate specificity, kinetic analysis	(Hardeman and Sjoling, 2007)
Lipase	Oil contaminated soil (Northern Germany)	*E. coli*/ cosmid	n.a	Agar-based assay- 1% tributyrin	30	7	Isolation of protein from inclusion bodies, refolding, protein purification, temperature, pH, effects of metals ions, solvent and various chemical, substrate specificity	(Elend *et al.,* 2007)
Lipase	Deep sea sediment of Edison Seamount	*E. coli*/ fosmid	1/8823	Agar-based assay-1% tricaprylin	25	8	Protein purification, temperature, pH, substrate specificity, effects of metal ions and detergent	(Jeon *et al.,* 2009b)
Lipase	Intertidal flat sediment	*E. coli*/ fosmid	1/6000	Agar-based assay - tributyrin	30	8	Protein purification, temperature, pH, effects metals ions and detergents, substrate specificity, conformational stability	(Kim *et al.,* 2009)

Table 1.1: *Continued from previous page.*

Enzyme	Environment	Host/ Vector	Positive Clones/ Total Clones	Screening Technique/ Substrate	T_{opt} (°C)	pH_{opt}	Level Of Characterization	Reference
Lipase	Soil from different altitude of Taishan (China)	*E. coli*/ plasmid	2/n.a	Agar-based assay-1% tributyrin	20	7 to 9	Protein purification, °C, pH, substrate specificity, effects of metal ions, kinetic analysis	(Wei *et al.,* 2009)
Lipase/ esterase	Activated sludge	*E. coli*/ plasmid	1/ 100000	Agar-based assay-1% tributyrin	10	7.5	Protein purification, temperature, pH, effects of detergents	(Roh and Villatte, 2008)
Lipase/ esterase	Antarctic soil	*E. coli*/ BAC	14/117742	Agar-based assay-1% tributyrin	35	basic	Enzymatic assay and effects of temperature using cell extract	(Berlemont *et al.,* 2011)
Esterase	Deep sea sediment (Papua New Guiney)	*E. coli*/ fosmid	1/n.a	Agar-based assay-1 % tributyrin	50-55 (high activation energy at 10-40°C)	10 to 11	Protein purification, temperature, pH, effects of metal ions and detergent, substrate specificity	(Park *et al.,* 2007))
Esterase	Antarctic desert soil	*E. coli*/ fosmid	3/ 100000	Agar-based assay- 1% tributyrin, 1% gum arabic,	40, (active at 7-54)	alkaline	Isolation of protein from inclusion bodies, refolding, protein purification, temperature, pH, substrate specificity	(Heath *et al.,* 2009)
Esterase	Arctic seashore sediment	*E. coli*/ fosmid	6/60132	Agar-based assay -1% tributyrin	30	8	Protein purification, temperature, pH, substrate specificity, enantioselective resolution of racemic ofloxacin esters	(Jeon *et al.,* 2009a)

Table 1.1: *Continued from previous page.*

Enzyme	Environment	Host/ Vector	Positive Clones/ total clones	Screening Technique/ Substrate	T_{opt} (°C)	pH_{opt}	Level Of Characterization	Reference
Esterase	Antarctic desert soil	*E. coli*/ fosmid	1/na	Agar-based assay 1% tributyrin	20	11	Isolation of protein from inclusion bodies, refolding, protein purification temperature, pH, substrate specificity, kinetic analysis	(Hu *et al.*, 2007)
Esterase	High Arctic intertidal zone sediment	*E. coli*/ fosmid	na	Agar-based assay- 1% trybutyrin	35	7.5	Protein purification, temperature and pH, substrate specificity, thermostability, crystallization and structutal analysis	(Fu *et al.*, 2012)
Amylase	Soil of Northwestern Himalayas	*E. coli*/ cosmid	1/ 350000	Agar-based assay- 1% starch; later stain with KI2	40	6.5	Protein purification, temperature & pH, effects of metal ions	(Sharma *et al.*, 2010)
Amylase	Antarctic soil	*E. coli*/ BAC	14/ 117742	Agar-based assay-0.5% starch, library later exposed to sublimated iodine vapour	35	basic	Enzymatic assay and effects of temperature using cell extract	(Berlemont *et al.*, 2011)
Cellulase	Antarctic soil	*E. coli*/ BAC	11/ 117742	Agar-based assay-0.5% Carboxymethylcellulose (CMC)	35-55	basic	Enzymatic assay and effects of temperature using cell extract	(Berlemont *et al.*, 2011)

Table 1.1: *Continued from previous page.*

Enzyme	Environment	Host/ Vector	Positive Clones/ total clones	Screening Technique/ Substrate	T_{opt} (°C)	pH_{opt}	Level Of Characterization	Reference
Cellulase	Antarctic soil	*E. coli*/ BAC	11/10000	Agar-based assay- 0.5% CMC and 0.01% Trypan blue	10 to 50	6 to 9	Protein purification, protein purification, temperature, pH, effects of various chemical, substrate specificity, viscometric assay	(Berlemont *et al.,* 2009)
β-glucosidase	Alkaline polluted soil, southern China	*E. coli*/ plasmid	2/30000	Agar-based assay- Esculin hydrate and ferric ammonium citrate	30 & 25	10	Protein purification, temperature, pH, thermostability, effects of inhibitors, kinetics	(Jiang *et al.,* 2011)
Protease	Antarctic soil	*E. coli*/ BAC	3/117742	Agar-based assay- 1% casein	na	na	Na	(Berlemont *et al.,* 2011)
Xylanase	Goat rument contents	*E. coli*/ plasmid		PCR amplification	30	6.5	Protein purification, temperature, pH, substrate specificity	(Wang *et al.,* 2011)
Alkane monooxygenase	Antarctic sediment	*E. coli*/ plasmid	177/na	PCR amplification	na	na	Gene sequencing	(Kuhn *et al.,* 2009)
DNA polymerase 1	Glacial ice (Germany)	*E. coli*/ plasmid and fosmid	15/23000	Complementation - growth assay	na	na	Subcloning into expression vector	(Simon *et al.,* 2009a)

1.4.1.1 Agar-based assays

One of the approaches in function-based metagenomic screening used to recover novel enzymes from low temperature habitat is by phenotypical detection of desired activity. As shown in Table 1.1, the most commonly reported type of functional screen is based on colony phenotype changes on agar plates. These assays typically rely on the utilization of a substrate, such as chemical dyes and insoluble or chromophore-bearing derivatives of enzyme substrates (Ferrer *et al.,* 2009), in the growth medium resulting in the appearance of zone around positive clones (Kennedy *et al.,* 2010).

Common substrates incorporated into agar to detect lipase or esterase activity include tributyrin, tricaprylin (Elend *et al.,* 2007, Jeon *et al.,* 2009b) and olive oil with Rhodamine B (Cieśliński *et al.,* 2009). For detecting proteases, skimmed milk is the most common substrate being used (Lammle *et al.,* 2007, Jones *et al.,* 2007, Waschkowitz *et al.,* 2009). Colonies that produce a halo on a purple background indicate an ability to use soluble starch in agar stained with iodine (Sharma *et al.,* 2010), while carboxymethyl cellulose alone or with addition of a specific dye is used to detect cellulase activity (Berlemont *et al.,* 2009). For example, screening of a metagenomic library derived from an Antarctic soil sample allowed the identification of a cold-adapted protein (RBcel1) that hydrolyses only carboxymethyl cellulose. This new enzyme displayed an endoglucanase activity, producing cellobiose and cellotriose, using carboxymethyl cellulose added with trypan blue as a substrate (Berlemont *et al.,* 2009).

1.4.1.2 Heterologous complementation

Another approach for detection of a desired enzyme is by heterologous complementation of host strains or mutants. For example by being able to select for activity rather than screening, the use of *E. coli* strains that require heterologous complementation for viability has been found to be an effective means for isolating genes with DNA polymerase 1 activity (Simon *et al.,* 2009a). Nine novel genes encoding complete DNA polymerase I proteins or domains typical of these proteins were isolated from metagenomic libraries constructed from glacial ice of the Northern Schneeferner, Germany and have potential use in molecular biology research (Simon *et al.,* 2009a).

1.4.1.3 Induced gene expression

In general, function-based detection of novel bioactive molecules is labour intensive, requiring screening of upwards of tens of thousands of clones. One process that can increase the throughput of screening metagenome libraries, is substrate-induced gene expression screening (SIGEX) (Uchiyama *et al.,* 2005). SIGEX is based on the premise that catabolic gene expression is generally induced by substrates and metabolites of catabolic enzymes and, in many cases, is controlled by regulatory elements situated in close proximity to catabolic genes. This approach employs an operon-trap *gfp*-expression vector that is suitable for shotgun sequencing and uses fluorescence-activated cell sorting (FACS) for high-throughput selection of positive clones in liquid cultures (Uchiyama *et al.,* 2005).

Uchiyama *et al*., introduced a related screening method, termed product-induced gene expression (PIGEX)(2010). The method is using a similar reporter assay-system, whereby the expression of the green fluorescent protein (GFP) is triggered by product formation. In response to benzoate production by the metagenomic clones, the sensor cells will show fluorescence signal (Uchiyama *et al.,* 2010). Three novel genes encoding amidases were identified from 96000 metagenomic clones derived from activated sludge using this method. As yet, no reported cold-adapted enzymes have been obtained by using this screening approach.

1.4.2 Challenges and improvements of functional screening

Although the attractiveness of functional screening lies in the potential of accessing novel enzymes and the guarantee to detect only enzymes that are active, it is compromised by the drawbacks of heterologous gene expression. In most function-based screening studies, the number of positive clones obtained have been very low (typically less than 0.01%) (Cowan *et al.,* 2005). Currently, *E. coli* is still the dominant screening host for functional metagenomics. Most of the commercially available large insert library production systems utilize *E. coli* as a replication host (Taupp *et al.,* 2011). Heterologous gene expression levels in *E. coli* are affected by several factors. These include, but are not limited to, plasmid copy number, mRNA stability, upstream elements, temperature and codon usage (Baneyx 1999b, Lithwick and Margalit 2003, Gustafsson *et al.,* 2004). This means that if the desirable bioactive molecules are not properly expressed they may not be detected.

Many approaches are being developed to mitigate this limitation. Improved systems for heterologous gene expression have been developed using shuttle vectors that facilitate screening of the metagenomic libraries in diverse host species. For example, there are developments of broad range shuttle vectors which can replicate in both Gram-positive and Gram-negative bacteria (Ono *et al.,* 2007, Troeschel *et al.,* 2012). Apart from that, low temperature expression systems have been developed by utilizing plasmids native to psychrophiles, including the Gram-negative Antarctic bacteria, *Psychrobacter* sp. (Tutino *et al.,* 2000) and *Pseudoalteromonas haloplanktis* (Tutino *et al.,* 2001). The origin of replication from the *P. haloplanktis* multicopy plasmid, pMtBL, was used to construct an *E. coli* shuttle vector utilizing a commercial pGEM plasmid (Tutino *et al.,* 2001). This shuttle vector was able to be stably maintained in five cold-adapted Gram-negative bacteria and was used to express a heat labile α-amylase in *P. haloplanktis* (Tutino *et al.,* 2001). Another low temperature expression system used cold-adapted *Shewanella* sp. Strain AC10 as host. Maximum recombinant protein expression was obtained when promoter for putative alkyl hydroperoxide reductase was used to prepare the construct of broad-host range vector pJRD215 (Miyake *et al.,* 2007).

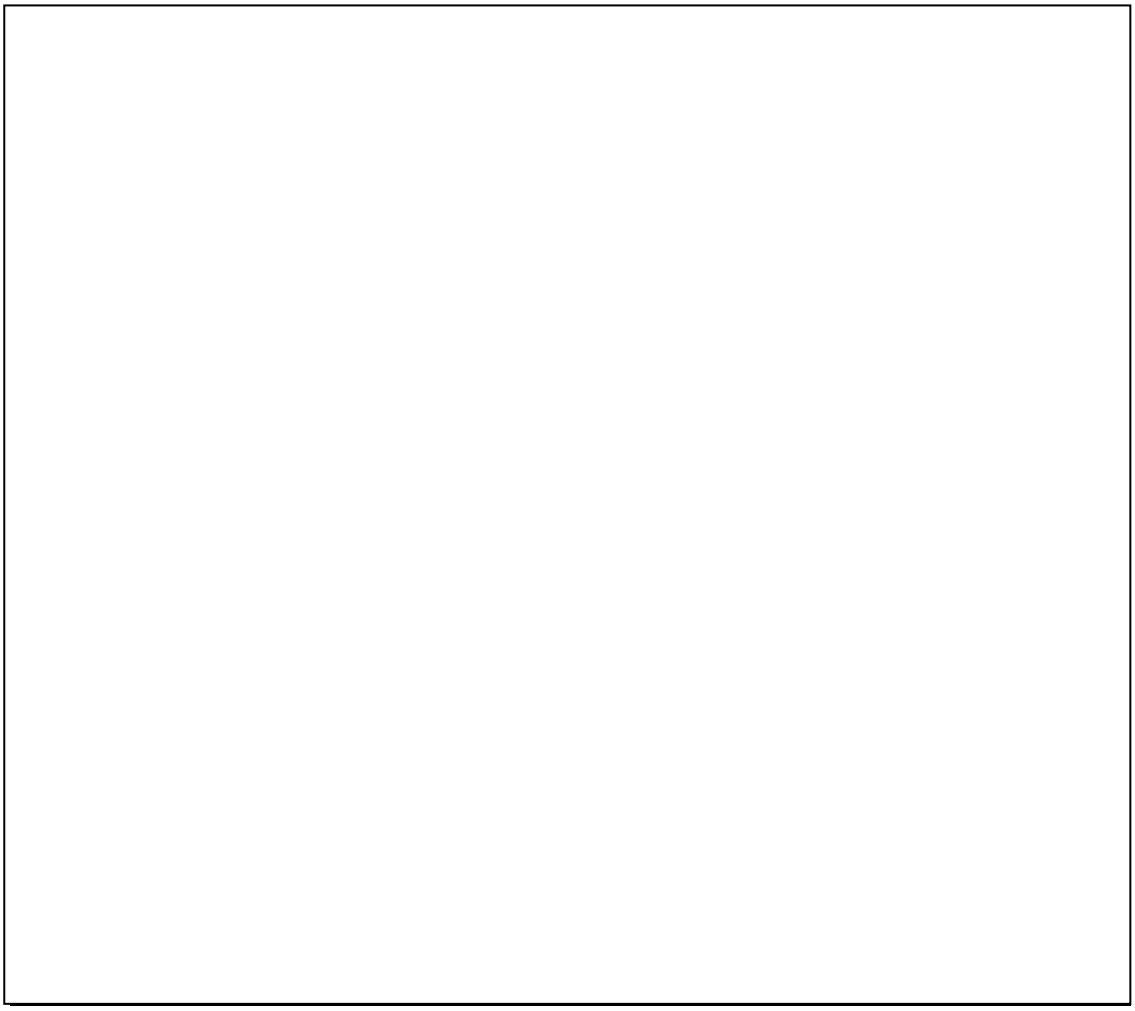
Protein expression systems that have been constructed for a cold-adapted microorganism as the host have an advantage, as it is possible to decrease the cultivation temperature to 0°C. This alleviates the heat denaturation of proteins and is suitable for the production of thermolabile proteins and reduces the possibility of inclusion bodies formation (Miyake *et al.,* 2007).

1.5 Sequence-based screening

The application of sequence-based approaches typically involves the design of DNA probes or primers which are derived from conserved regions of already known genes or protein families (Simon and Daniel, 2009). As this approach is dependent on established sequence databases, it is limited in its ability to detect novel gene families. Nevertheless, molecular diversity is so great that surprisingly novel enzyme sequences have been retrieved using this approach, such as a novel alkane monooxygenase (Kuhn *et al.,* 2009) and xylanase (Wang *et al.,* 2011). Unlike function-based methods, sequence-based detection of target genes is possible, regardless of gene expression and protein folding in the host, and irrespective of the completeness of the target gene's sequence (Lorenz *et al.,* 2002). The success of this approach depends on the advances in high-throughput, bioinformatic tools for metagenomic analysis and sequence databases supported by experimental data, and the lowering cost of sequencing technologies.

DNA sequencing technologies can be classified into several categories (Table 1.2) (Scholz *et al.,* 2012, McGinn and Gut, 2013). Second generation DNA sequencing technology is the current state of the art and is quickly becoming the standard method for sequencing. These high-throughput sequencing technologies (e.g. 454 and Illumina) can potentially generate 10^6–10^9 sequences (100–700 bp) per run (Scholz *et al.,* 2012). Such high read depths yields greatly improved coverage of species within the community (Metzker, 2010). However, one limitation is that the generally shorter fragment lengths do not contain the full open reading frame (ORF) for the functions of interest (Li *et al.,* 2009).

Table 1.2**:** List of new generation sequencing platforms and their expected throughputs, error types and error rates. Each platform has distinct advantages owing to cost, error rate, read length and so on. Adapted from (Scholz *et al.,* 2012). Table has been removed due to Copyright restrictions.



There are several publications that describe searching metagenomic sequence databases directly for genes that encode for potential commercially useful enzymes (Schlüter *et al.,* 2008, Yergeau *et al.,* 2010, Cavicchioli *et al.,* 2011). Screening can be performed by searching for primary sequence identity and motifs, and by evaluating predicted protein structures and putative catalytic sites that match to known enzymes (Cavicchioli *et al.,* 2011). For example, one analysis of Arctic permafrost metagenome data identified trehalase, chitinase, β-glucosidase and β-galactosidase genes (Yergeau *et al.,* 2010). In another report, a metagenome library of an agricultural biogas fermenter community sequenced using 454-pyrosequencing technology identified gene regions with cellulolytic functions matching to a *Clostridium* genome, which indicates this genus is important for hydrolysis of cellulosic plant biomass (Schlüter *et al.,* 2008).

The next revolution in sequencing known as the 'third generation'-PacBio RS by Pacific Biosciences is already underway offering longer reads and shorter run times (Logares *et al.,* 2012). New bioinformatics tools are being developed to assist in efficient data mining, based not only on primary sequence homology but also on protein structure, catalytic sites, and activity prediction (Roy *et al.,* 2010, Peng *et al.,* 2011). Improvements in the speed and cost of *de novo* gene synthesis have facilitated the complete redesign of entire gene sequences to maximize the likelihood of high protein expression (Gustafsson *et al.,* 2004). It is anticipated that with the advancement of DNA sequencing techniques, soon sequence-based metagenomic databases searches, combined with bioinformatic tools as well as commercial gene synthesis will have a greater influence on mining novel biocatalyst genes than function-based methods.

1.6 Ace Lake: Potential resources of cold-adapted enzymes

The greatest concentration of stratified water bodies in Antarctica, and possibly the world, is found in the Vestfold Hills, where both meromictic (permanently stratified) lakes and stratified marine basins occur (Burton, 1981, Burke and Burton, 1988, Gibson, 1999). The Vestfold Hills is an ice free oasis which lies on the eastern side of Prydz Bay (Gibson, 1999). The retreat of the continental ice-shelf lead to isostatic uplift of the land and resulted in the formation of freshwater lakes in the area closer to the ice sheet (Laybourne-Parry and Marchant, 1992) and saline or hypersaline lakes in the area closer the coastline (Burton, 1981).

Ace lake (68.473°S, 78.189°E) is one of the pristine, meromictic lakes located on the Long Peninsula in the Vestfold Hills (Figure 1.2) (Rankin *et al.,* 1999). During the early Holocene (13000-9400 years ago), Ace Lake was an aerobic freshwater system that evolved to an open marine basin influenced by dynamic mixing of ocean and meltwater inputs. Marine input into the lake ceased approximately 5500 years ago and Ace Lake became a meromictic basin which stabilised to become the lake that it is today with little change for about the past 4000 years (Fulford-Smith and Sikes, 1996). It has an area of 18 h.a, a maximum depth of 25 m and a salinity range from approximately 20 to 40 gL^{-1} (Lauro *et al.,* 2011). Vestfold Hills experiences below freezing temperatures for most of the year resulting in Ace Lake having an ice cover for approximately eleven months of the year (Rankin *et al.,* 1999). The temperature of the lake water is in the range 0-3.5°C (Lauro *et al.,* 2011).

The organisms now living in the lake are marine-derived and therefore Ace Lake represents a unique model system for studying the evolution of marine microbiota (Fulford-Smith and Sikes, 1996). Microbiota have been sampled from the lake water column at different zones, each zones differs in temperature, oxygen, salinity and other physico-chemical parameters (Lauro *et al.,* 2011). Unusual microbial signatures, such as isolates of novel fermentative "coiled bacterium", a cell wall-less member of *Spirocheteae* and psychrophilic methanogens, *Methanogenium frigidum* and *Methanococcoides burtonii* (Franzmann *et al.,* 1997, Gibson, 1999, Rankin *et al.,* 1999), have already been identified (Franzmann and Rohde, 1991, Franzmann and Dobson, 1992, Allen *et al.,* 2009).

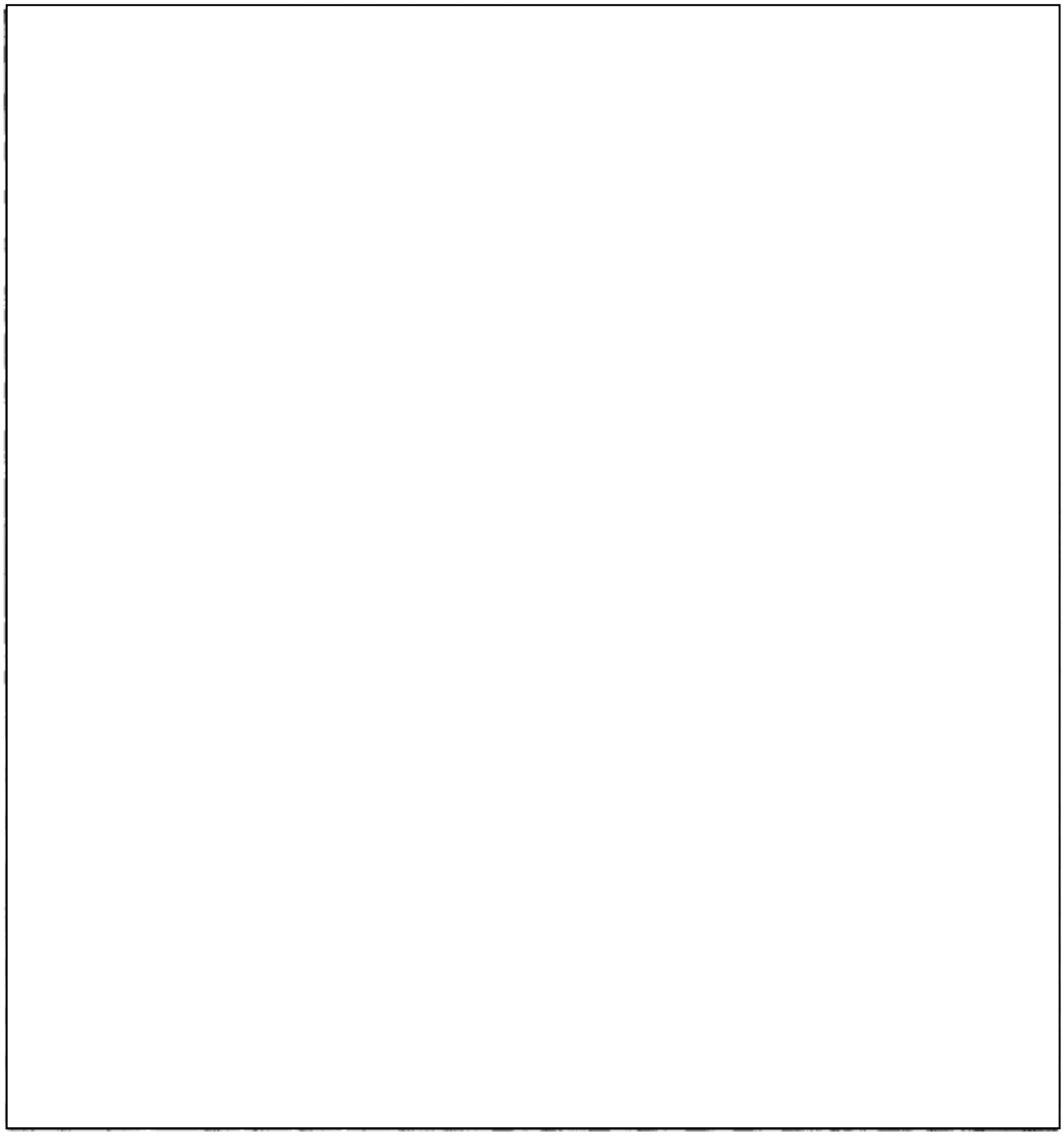


Figure 1.2: A map of the Vestfold Hills showing fjords, bays and lakes (numbered). The lakes are: (1) unnamed lake 2, (2) Organic Lake, (3) Pendant Lake, (4) Glider Lake, (5) Ace Lake, (6) unnamed lake 1, (7) Williams Lake, (8) Abraxas Lake, (9) Johnstone Lake, (10) Ekho Lake, (11) Lake Farrell, (12) Shield Lake, (13) Oval Lake, (14) Ephyra Lake, (15) Scale Lake, (16) Lake Anderson, (17) Oblong Lake, (18) Lake McCallum, (19) Clear Lake, (20) Laternula Lake (21) South Angle Lake, (22) Bayly Bay, (23) Lake Fletcher, (24) Franzmann Lake, (25) Deprez Basin, (26) 'Small Meromictic Basin', Ellis Fjord, (27) Burton Lake, (28) Burch Lake, (29) Tassie Lake, (30) Club Lake, (31) Lake Jabs, (32) Deep lake, (33) Lake Stinear, (34) Lake Dingle, (35) Lake Druzhby, (36) Watts Lake, (37) Lebed Lake and (38) Crooked Lake. Map image is from (Gibson, 1999) with some minor modification**.** Figure has been removed due to Copyright restrictions.

Recent metagenomic analysis confirmed that the cellular microorganisms dominating Ace Lake are bacteria, with relatively few eucarya present, and few archaea in the anaerobic zones (Lauro *et al.,* 2011). Large algal viruses (*Phycodnaviridae*) were also detected in the lake. The nature of this habitat, its permanently cold environment, alkaline pH, salinity and its microbial diversity lead itself to the discovery of potentially unique cold-adapted enzymes (Laybourn-Parry and Pearce, 2007). Enzymes with activity at alkaline pH have many potential applications such as detergent additives in the detergent industry, hide-dehairing process in leather industry, alkali-treated wood pulp bleaching process in pulp and paper industry and industrial production of cyclodextrin for foodstuffs, chemicals, and pharmaceutical use (Horikoshi, 1999).

1.7 Objectives

Previous studies have provided important information relative to the nature of cultured and uncultured microbial community and their functional potential in the Ace Lake aquatic environment. Verification of the functional potential through the expression study should be a meaningful way of exploitation of the natural resources from this cold-adapted environment. The overall aim of this study was to use the existing metagenomic resources from the Antarctica, particularly from the Ace Lake, Organic Lake and Southern Ocean aquatic environment, to identify and analyse the genes coding for hydrolases, to screen for enzymatic activity in the metagenomic clones through agar-based assays and to manipulate the identified genes in the overexpression studies. The specific aims of this study were described in four chapters (Chapter 2, 3, 4 and 5) and concluded with general discussion and the future perspectives in Chapter 6.

Chapter 2 describes the sequence-based screening of translated protein sequences in the Ace Lake metagenomic dataset for hydrolases. The screening was performed using specific profiles Hidden Markov Model (HMM) for subtilase, lipases and glycosyl hydrolases. The datasets utilized in the screening processes consisted of Sanger sequencing reads, assemblies of Sanger and 454 pyrosequencing reads and assemblies of 454 pyrosequencing reads. The results were analysed according to the E-value cut-off, protein length, and annotation from the public database including Kyoto Encyclopedia of Genes and Genomes (KEGG), NCBI-nr and Swiss-Prot databases. Phylogenetic analysis was performed to determine the diversity of

the genes identified from the sequence-based screening. The possibility of functional novelty of the Ace Lake derived-sequences was also explored through domain architecture analysis.

In Chapter 3, the screening of the Ace Lake metagenomic clones for protease, lipase, and amylase on agar-based assays is described. The clones utilized for the screening purposes are based on two groups. The first group was the targeted clones which contained the specific subtilase, lipase and glycosyl hydrolase genes that were previously identified through sequence-based screening in the Chapter 2. These clones were retrieved from the clone library through barcode identification. The second group consisted of ~20000 randomly-picked clones. Any activity detected on agar-based assays was subsequently confirmed by liquid enzymatic assay.

Chapter 4 describes the expression studies of three subtilase genes (Subt9195, Subt8715 and Subt5372) in pET expression system. These genes were identified by the analysis conducted in Chapter 2 and isolated from the Ace Lake metagenomic clone library for further manipulations. At least two approaches were taken to solubilise the protein that was expressed as inclusion bodies. The first approach was to solubilise the inclusion bodies in a range of denaturants and solubilising agents and refold the solubilised protein in a range of refolding buffers. The second approach was to change the expression vector. This new construct included the sequence for the high molecular weight tag, NusA, which is known to increase solubility of the recombinant proteins during expression. The vector also included the sequence of His-tag for His-tag protein purification of the solubilised protein. The identity of the solubilized protein was identified by fourier transform mass spectrometry (FTMS). Protease activity was determined using a skimmed milk agar-based assay and a liquid assay using ρ-N-succinyl-AAPF-ρ-nitroanilide as the substrate. Structure modelling of the expressed proteins was included based on the available bioinformatic data, since the purification and protein crystallization were beyond the scope of this thesis.

In Chapter 5, the relative abundance of the subtilase and other peptidase genes in the uncultivated microbial community in the Ace Lake aquatic environment and neighbouring sites, i.e. Organic Lake and the Southern Ocean was investigated. This was part of a larger study to understand the roles of diverse microorganisms in cold-adapted environment and their biotechnological potential. Subtilase and other peptidase sequences were analysed

according to Clusters of Orthologous Groups (COGs) categories and subjected to appropriate statistical test. Through KEGG annotations, the associated taxonomic diversities were analysed to infer the physiological roles of the peptidases in the three different cold-adapted environments.

Chapter 6 summarises all the findings that are reported in this thesis. The chapter also consider the future work for bioprospecting of enzymes from the Antarctic metagenome. It was concluded that the studies presented in this thesis establish a foundation for future bioprospecting of enzymes from the Antarctic aquatic environment.

CHAPTER 2

Sequence-based screening of hydrolases in the Ace Lake metagenomic datasets using profile Hidden Markov Models

2.1 Introduction

2.1.1 Metagenomics and microbial life in cold environments

The microorganisms thriving in permanently cold environments, such as in the polar regions, have adopted a variety of adaptive strategies to maintain activity and metabolic function to face the conditions considered harsh from an anthropocentric point of view. These physiological challenges include persistently low temperatures, low abundance of nutrients, as well as frequent freeze–thaw cycles, salinity fluctuations, desiccation and varying seasonal light conditions (Deming, 2009, Margesin and Miteva, 2011). Cold adaptation occurs at both the cellular and molecular levels including alterations in membrane fluidity, expression of cold-shock and cold acclimation proteins that regulate transcription and translational processes, antifreeze/ice-nucleating proteins, production of compatible solutes and exopolysaccharides and enzymes that are capable of catalysing chemical reactions at low temperature (D'Amico *et al.,* 2006, Siddiqui and Cavicchioli, 2006).

Metagenomics, a direct analysis of genes contained in an environmental sample, has enabled a culture-independent assessment of microbial communities in their environment (Simon and Daniel, 2009, Casanueva *et al*. 2010). This technique offers insight into the diversity and distribution of cold-adapted microorganisms, prediction of the functional roles of the organisms detected, as well as being an invaluable resource for discovery of novel enzymes with unusual properties (Ferrer *et al.,* 2009, Cavicchioli *et al.,* 2011). Metagenomics-based studies in permanently cold environments has revealed that microorganisms have numerous metabolic adaptations associated with a psychrophilic lifestyle and a broad range of decomposition and nutrient recycling potentials (Casanueva *et al.,* 2010, ,Yergeau *et al.,* 2010, Varin *et al.,* 2012). However, the adaptation of psychrophiles to the extreme conditions of their habitat means they require specialized temperature controlled equipment (and associated energy costs for operation) to enable

their growth in the lab (Hoag, 2009; Cavicchioli *et al.* 2011) which may be difficult to replicate. This realization has further strengthened the need to employ cultivation-independent approaches to allow comprehensive access to the genetic diversity of microbial communities for bioprospecting purposes (de Pascale *et al.,* 2012).

2.1.2 Metagenomics of Ace Lake

The Cavicchioli research group has successfully performed a metagenomic study of a pristine, marine-derived, stratified lake (Ace Lake) in Antarctica (Lauro, *et al.* 2011). As illustrated in Figure 2.1, Ace Lake can be divided into three main zones: an upper-oxic mixolimnion that extends to 11.5 m, a transition zone corresponding to the halo- and oxycline centered at approximately 12.7 m depth and a lower anoxic monimolimnion (Rankin *et al.,* 1999, Lauro *et al.,* 2011). In the study, nine million ORFs were analysed; representing microbial samples taken from the six depths of the lake (5, 11.5, 12.7, 14, 18 and 23 m) and size fractionated sequentially onto 3.0, 0.8 and 0.1 μm filters. The sequences were obtained using a combination of Sanger sequencing and 454 pyrosequencing technologies (Lauro *et al.,* 2011).

The phylogenetic analyses of the metagenome indicated that the cellular microorganisms dominating the lake are bacteria, with relatively few eucarya present (mainly in the mixolimnion), and few archaea in the monimolimnion. The study inferred that *Flavobacteria* and *Gammaproteobacteria* are the bacterial members in the mixolimnion, responsible for remineralisation of particulate organic carbon to dissolved organic carbon (DOC). Free-living, oligotrophic *Actinobacteria* and members of the SAR11 clade would further perform heterotrophic conversion of DOC. In the monimolimnion, the combination of fermentative, sulfate-reducing and methanogenic microorganisms would decompose particulate organic carbon into smaller molecules, ultimately to CO_2 and CH_4 (Lauro *et al.,* 2011).

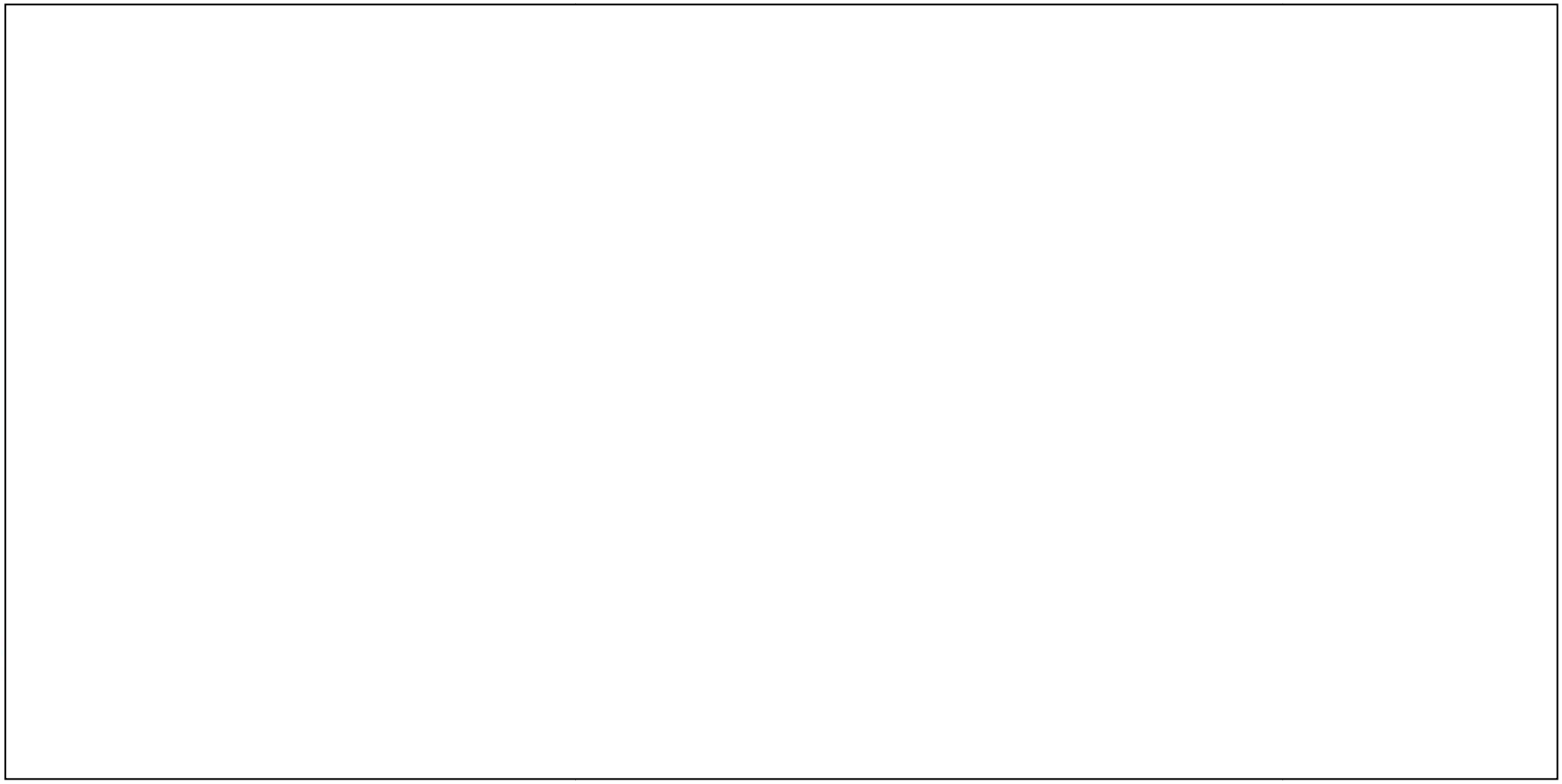


Figure 2.1**:** Physico-chemical and biological structure of the Ace Lake. The figure is from (Lauro *et al.,* 2011) with some minor modifications. Figure has been removed due to Copyright restrictions.

2.1.3 Sequence-based screening of the metagenomic data using profile Hidden Markov Models

Proteins can be classified into families of related sequences and structures (Henikoff *et al.,* 1997). Multiple sequence alignments of sequences from similar protein families can reveal patterns of conservation. Profile Hidden Markov Models (HMM) are statistical models of multiple sequence alignments. The HMM identifies a set of positions that described the conserved primary structure from a given family of proteins, or in other words, the core elements of homologous protein sequences (Krogh *et al.,* 1994).

The advantage of using profile HMM is it has a consistent theory for setting position-specific gap and insertion scores thus make it highly automatable and fast. This method has allowed the construction of libraries of hundreds of profile HMMs and the application on a very large scale to whole genome analysis. One of the profile HMMs-based protein domain libraries is Pfam database (Finn *et al.,* 2010). This comprehensive database of over 12000 conserved protein domain families is widely used by biologists to annotate and classify proteins (Finn *et al.,* 2010).

2.1.4 Classification of hydrolases based on sequence homology

2.1.4.1 Subtilase: The superfamily of subtilisin-like proteases

The Subtilase Family, also called Family S8 in the comprehensive peptidase database (MEROPS) classification, is the second largest family of serine peptidases, both in terms of number of sequences and characterized peptidases (Rawlings *et al.,* 2006). The overall sequence identity of the protease domain throughout the entire family is low except for the short segments surrounding the Asp-His-Ser catalytic site, and the Asn residue, which contributes to the oxyanion binding site (Siezen and Leunissen, 1997, Siezen *et al.,* 2007). Based on sequence similarity of the catalytic domain, subtilases are divided into six families: subtilisin, thermitase, proteinase K, lantibiotic peptidase, kexin and pyrolysin (Siezen and Leunissen 1997).

Subtilases are widely distributed in various organisms including archaea, bacteria, viruses, fungi, yeast and higher eukaryotes (Siezen and Leunissen, 1997). In bacteria, archaea and fungi, most of the subtilases are secreted and involved in the degradation of protein for amino acid uptake. The enzymes have an important role in biotechnology. For example, proteinase K is used as reagent to make peptides from proteins and subtilisin is widely used as an active agent in detergent (Rawlings *et al.,* 2006).

2.1.4.2 Glycoside Hydrolases

Glycoside hydrolases (GH)s are a widespread group of enzymes that hydrolyse the glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety (Reddy *et al.,* 2004). These enzymes are classified into more than 100 families based on sequence homology, which is available through the Carbohydrate Active enZyme database (http://www.cazy.org/)(Cantarel *et al.,* 2009). Cellulase is grouped in glycoside hydrolase family 5 (GH5) while α-amylase is grouped in glycoside hydrolase family 13 (GH13). Glycoside hydrolase family 16 includes lichenase and agarase. Enzymes of glycoside hydrolase family 20 (GH20) cleave β-1,4-linked N-acetylglucosamine and β-1,6-linked N-acetylglucosamine residues, have potential as an antibiofilm agent (Kerrigan *et al.,* 2008).

Most of the known starch-modifying enzymes can be found in the GH13 group (MacGregor *et al.,* 2001). Enzymes in this group were also known as the α-amylase family, which includes α-amylase, α-1,6-glucosidase, branching and debranching enzymes, maltogenic amylase, neopullulanase, trehalose synthase and cyclodextrinase (CDases) (Horvathova *et al.,* 2001, MacGregor *et al.,* 2001). GH13 is the largest sequence-based family of glycoside hydrolases acting on starch, glycogen, and related oligo- and polysaccharides (Stam *et al.,* 2006). The complexity of this family has driven numerous analyses attempt to derive relationships between the sequence and the properties of the enzymes (Jespersen *et al.,* 1993, Janecek *et al.,* 1997, MacGregor *et al.,* 2001). There are seven conserved regions in the GH13 enzymes which are useful for identification of the specific enzyme group (Janeček, 2002). For example, the oligo-1,6-glucosidase group has the template sequence QPDLN in the fifth conserved region compared to α-amylase that has a conserved LPDLD sequence (Janeček, 2002).

2.1.4.3 Lipases

Lipases (triacylglycerol acylhydrolase, EC 3.1.1.3) are part of the family of hydrolases that act on carboxylic ester bonds. In addition to this natural function, lipases can catalyse esterification, inesterification, and transesterification reactions in non-aqueous media (Houde *et al.,* 2004). Lipases are widely distributed in nature but are known to be abundant in bacteria, fungi and yeast (Pandey *et al.,* 1999). Microbial lipases display wide substrate specificity, a property that seems to have evolved to ensure the access of the microorganisms to diverse carbon sources during the recycling of lipid-containing nutrients (Gunstone, 1999, Bornscheuer, 2002). Based on conserved sequence motifs and biological properties, lipases are grouped into eight families (Arpigny and Jaeger, 1999). Several extensions of the original classification scheme such as the enlargement of family

6 to 7 subfamilies was proposed to include the more recently discovered lipase types (Jaeger and Eggert, 2002).

Family I, also known as true lipases is comprised of lipase similar to *Pseudomonas* lipases that were probably the first to be studied and have various industrial applications (Arpigny and Jaeger, 1999). α/β hydrolases notably include lipolytic enzymes contain Ser-Asp-His as their catalytic site. For true lipases, the serine residue usually appears in the conserved pentapeptide Gly-Xaa-Ser-Xaa-Gly (Arpigny and Jaeger, 1999). Family II of the lipolytic enzyme (also known as lipase GDSL), possess a different GDSL sequence motif. GDSL enzymes have five conserved sequence blocks (I-V) and four invariant important catalytic residues Ser-Gly-Asn and His in blocks I,II,III and V respectively (Akoh *et al.,* 2004). The GDSL family is further classified as SGNH hydrolase because of the strict conservation of residues Ser-Gly-Asn-His in the conserved blocks I, II, III, and V (Chepyshko *et al.,* 2012). The other lipase families (III-VIII), which includes hormone sensitive lipase (HSL) and many esterases, are outside the scope of this work.

This chapter described the use of profile HMM-based sequence search, a sensitive, statistically sound analysis method capable of identifying remote homologs, to search for hydrolases sequences in the Ace Lake metagenomic dataset. The search was performed on the translated ORFs of Sanger reads, assemblies of 454 pyrosequencing reads and assemblies of combination of Sanger and 454 pyrosequencing reads.

2.2 Materials and methods

2.2.1 Ace Lake samples

The Ace Lake water samples were collected as described in (Lauro *et al.,* 2011) from Ace Lake (68° 24'S, 78° 11'E), Vestfold Hills, Antarctica on the 21 and 22 December 2006. Water samples from six depths of the lake, i.e., 5 and 11.5 m (mixolimnion), 12.7 m (interphase) and 14, 18 and 23 m (monimolimnion) were passed through a 20 μm pore size pre-filter, and microbial biomass was captured by sequential filtration onto 3.0 μm, 0.8 μm and 0.1 μm pore size 293 mm polyethersulfone membrane filters. DNA was extracted from the filters and samples were sequenced using the Roche GS-FLX titanium sequencer (454 pyrosequencing). Samples of the 0.1 μm were also sequenced using Sanger sequencing technology.

2.2.2 Dataset description

The Ace Lake metagenomic dataset derived from two sequencing technologies, i.e, Sanger sequencing and 454 pyrosequencing, were analysed in this chapter. There were two data samples from 0.1 μm. One sample was the combination of Sanger and 454 pyrosequencing reads and designated as the hybrid. Another sample was from Sanger sequencing reads. Data samples of 0.8 and 3.0 μm were from 454 pyrosequencing reads. ORFs were predicted using MetaGene (Noguchi *et al.,* 2006) from the reads of Sanger sequencing, the assembled reads of 454 pyrosequencing and the hybrids before subsequently translated. The summary of the data used throughout the analysis in this chapter is provided according to the sample ID, sampling depth (m), filter size applied (μm) and the total number of translated ORFs for both assemblies and reads (Table 2.1). Ace Lake is a meromictic lake which means that it has layers of water that do not intermix. As illustrated by Figure 2.1, the range of pH and salinity of each layer is relatively different. Sequence data from microbial biomass at different depth provide access to the diversity of microorganism throughout the lake.

Table 2.1: Summaryof the Ace Lake metagenomic dataset used in the sequence-based screening.

Sample ID	Water Depth (M)	Filter Size (μm)	Total Number Of Translated ORFs	
			Assemblies	Sanger Reads
232	5	0.1	138208*	484044
232	5	0.8	63959	-
232	5	3.0	28931	-
231	11.5	0.1	133948*	487078
231	11.5	0.8	37475	-
231	11.5	3.0	63746	-
230	12.7	0.1	27142*	85935
230	12.7	0.8	33933	-
230	12.7	3.0	36804	-
229	14	0.1	62436*	16793
229	14	0.8	81695	-
229	14	3.0	55936	-
228	18	0.1	71512*	17002
228	18	0.8	111344	-
228	18	3.0	50077	-
227	23	0.1	128878*	185761
227	23	0.8	110302	-
227	23	3.0	48384	-

*Assemblies prepared from combination of Sanger reads and 454 pyrosequencing reads (hybrid)

2.2.3 HMMsearch for hydrolases from Ace Lake metagenome dataset

The screening was performed on the translated ORFs from the Ace Lake metagenomic dataset listed in the Table 2.1. HMMer (Version 2.3.2) (http://hmmer.janelia.org/) was used to search for the homolog of the desired protein sequences using profile HMM that were retrieved from the HMM Pfam library. The profile HMM were selected based on the keyword search for subtilase, lipases and glycoside hydrolases in the Pfam database (Finn *et al.,* 2008). The initial rounds of the searches were performed using a default E-value cut-off of 10. All the selected profile HMM (listed in Table 2.2) have the identified set of positions of conserved sequences for each enzyme. The selected hydrolases that included peptidase S8-Subtilase Family, lipase 1, lipase 2, lipase 3, lipase GDSL, GH5, GH13, GH16 and GH20 have wide range of biotechnological applications. It is anticipated that the newly identified sequences encoded for cold-adapted enzymes with unique characteristics, such as the ability to be active at alkaline pH and saline environment.

Table 2.2: List of the selected profile HMM utilized in the sequence-based screening process.

PFAM ID	Profile HMM description
PF00082	PeptidaseS8- Subtilase Family
PF00151	Lipase 1
PF01674	Lipase 2
PF01764	Lipase3
PF00657	LipaseGDSL
PF00128	α-amylase(GH13)
PF00150	Cellulase (GH 5)
PF00722	GH16
PF00728	GH20

2.2.4 Selection of the sequences from the HMMsearch

To further test the robustness of the HMMsearch, matches at three E-value thresholds (1E+01, 1E-5 and 1E-20) were considered. The results at all three thresholds were identified and divided into two groups based on the sequence length and the catalytic sites. Sequences that have more than 200 amino acid (aa) residues long and have at least one catalytic site were grouped together in the 'lenient' group. The 'stringent' group consists of the sequences that have more than 200aa (subtilase and lipase) or 300aa (glycoside hydrolase), have all three catalytic sites and were predicted to be a complete ORFs by MetaGene (Noguchi *et al.,* 2006). Identification of the catalytic sites were based on the information obtained in the HMMsearch and multiple sequence alignment of the matches using MUSCLE (Edgar, 2004) with reference protein sequences (Appendix B).

2.2.5 Annotation of putative hydrolase sequences and domain architecture analysis

Annotation of the sequences in the 'stringent group' derived from the assemblies were retrieved from the in-house pipeline described by DeMaere *et al.* (2011). Sequence derived from the Sanger reads, were annotated manually by basic local alignment search tool (BLAST) comparison to NCBI-nr, SwissProt and KEGG-peptide sequence database. All sequences were further analysed in MEROPS (Rawlings *et al.,* 2006, Rawlings *et al.,* 2012) (peptidase only), Conserved Domain Database (CDD) (Marchler-Bauer *et al.,* 2009) and Conserved Domain Architecture Retrieval Tool (CDart) (Geer *et al.,* 2002). The results are summarized in the annotation table (Appendix C).

2.2.6 Phylogenetic analysis

For phylogenetic analysis, multiple sequence alignments were performed using MUSCLE (Edgar, 2004) and viewed with GeneDoc (http://www.psc.edu/biomed/genedoc). Phylogenetic trees were constructed in MEGA5 (Tamura *et al.,* 2011) using the Neighbor-Joining method. One thousands bootstrap replicates were used. The evolutionary distances were computed using the p-distance method.

2.3 Results

2.3.1 Distribution of the HMMsearch result for subtilase, lipases and glycoside hydrolases

The initial screen using the default E-value cut-off of 10, identified a total of 2683 genes with hydrolase domains in the Ace Lake metagenome dataset. The HMMsearch of four lipase groups (lipase 1, lipase 2, lipase 3 and lipase GDSL) from the Ace Lake metagenomic dataset has resulted with matches only to the two of the lipase groups; lipase 3 and lipase GDSL. Irrespective of the sampling depth and filter size, the breakdown of the types of hydrolases matches from the highest to the lowest count was α-amylase (778), subtilase (764), lipase GDSL (566), lipase 3 (310), cellulase (141), glycosyl hydrolase 16 (108) and glycosyl hydrolase 20 (96)(Table 2.3). Further analysis of the distribution of subtilase, lipases and GH13 according to the type of dataset (Sanger reads, assemblies of 454 pyrosequencing reads, and Sanger-454 pyrosequencing hybrid), sampling depths and filter size were illustrated in Figure 2.2 and Figure 2.3. The distributions of the matches in the HMMsearch results were described as the percentage number of counts relative to the number of translated ORFs in each sample. Further analysis to cellulase, glycosyl hydrolase 16 and 20 were not included in this chapter due

to low number of matches and result from MetaGene prediction indicated no complete ORF sequence.

In general, based on the E-value cut-off of 10, the relative percentage of matches for amylase and subtilase were higher compared to lipase 3 and lipase GDSL. The most obvious difference in terms of distribution of amylase and subtilase was at depth 12.7 m (interphase) where amylase showed the highest relative percentage of matches in at least three datasets, i.e., assemblies of Sanger-454 hybrid (0.1 μm), assemblies of 454 (3.0 μm) and Sanger sequences (0.1 μm)(Figure 2.2). Lipase GDSL matches was 1.8 fold higher compared to lipase 3 and showed the highest relative percentage of matches in the 11.5 m of assemblies of 454 (0.8 μm). Lipase 3 that generally showed lower relative percentage of matches compared to lipase GDSL, was higher in the 12.5 and 14 m sample of assemblies of Sanger-454 hybrid (0.1 μm), 5 m sample of assemblies of 454 (0.8 μm), 5 m sample of assemblies of 454 (3.0 μm) and in the 12.5, 14 and 18 m sample of Sanger sequences (Figure 2.3). When E-value cut-off of 1E-20 was applied, the matches for subtilase and amylase reduced to lower than 0.02% except in the 11.5 m of the assemblies of 454 (0.8 μm) and 12.7 m of the Sanger sequences. As for lipase 3 and lipase GDSL, the matches reduced drastically in E-value cut-off of 1E-20.

Table 2.3: The breakdown of the types of hydrolases identified in the Ace Lake metagenomic dataset with default E-value cut-off of 10.

Profile HMM description	**Total number of matches**
GH13	778
PeptidaseS8- SubtilaseFamily	764
LipaseGDSL	486
Lipase3	310
Cellulase	141
Glycosyl hydrolase 16	108
Glycosyl hydrolase 20	96

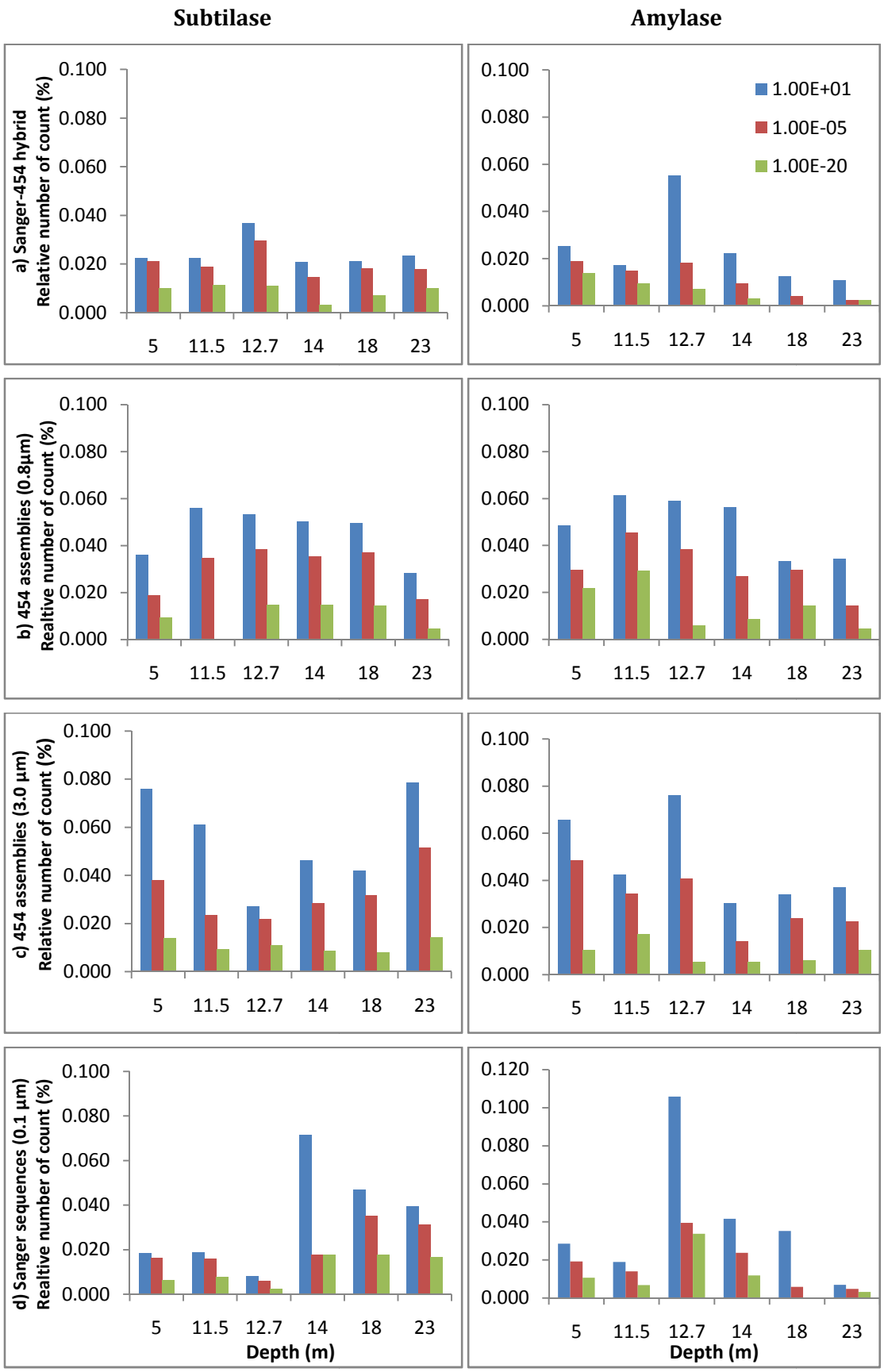


Figure 2.2: Distribution of subtilase (left) and amylase (right) in percentage number of counts relative to the number of translated ORFs in the metagenome datasets of (a) Sanger-454 hybrid (0.1 µm) (b) 454 assemblies (0.8 µm), (c) 454 assemblies(3.0 µm) and (d) Sanger sequences (0.1 µm).

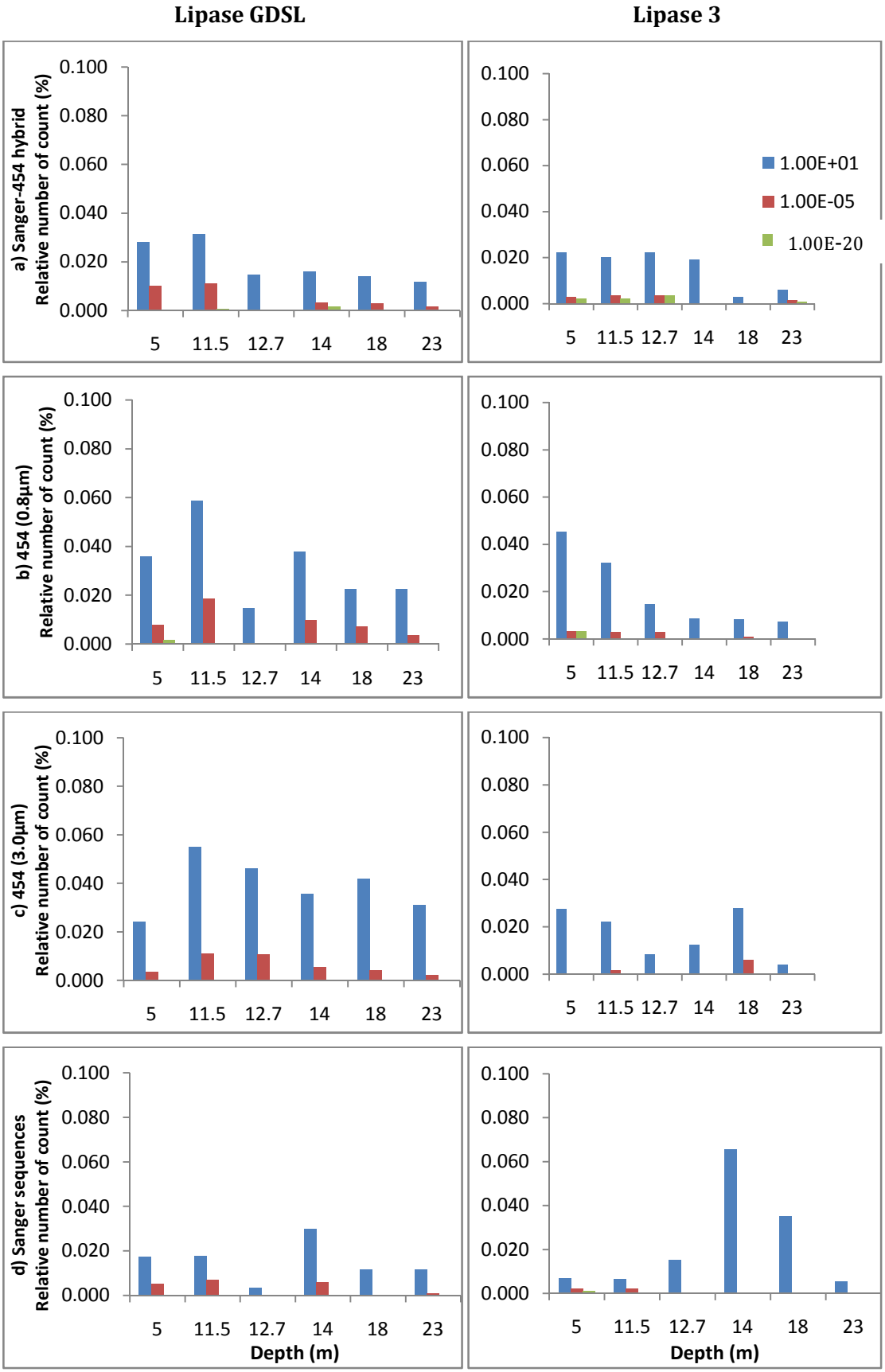


Figure 2.3: Distribution of lipase GDSL (left) and lipase 3 (right) in percentage numbers of counts relative to the number of translated ORFs in the metagenome datasets of (a) Sanger-454 hybrid (0.1 µm), (b) 454 assemblies (0.8 µm), (c)454 assemblies (3.0 µm) and (d) Sanger sequences (0.1 µm).

2.3.2 Selection of the subtilase sequences from the HMMsearch results

The matches to subtilase were further analysed based on their sequence length and E-value. Scatter plots of the number of matches in the 0.1 μm sample of Sanger-454 hybrid and Sanger reads from the three main zones of the Ace Lake (mixolimnion (5 and 11 m), interphase (12.7 m) and monimolimnion (14, 18, 23 m)) are shown (Figure 2.4). In comparison to the matches in the Sanger reads, the plots indicated more hydrolase matches in the Sanger-454 pyrosequencing hybrid consisted of longer translated ORFs with low E-value. This was apparent in the mixolimnion, where the longest translated ORF was up to 1200 aa. In contrast, most of the proteins predicted from Sanger reads were shorter (less than 300 aa). Short reads were all predicted to be partial ORFs.

Following a more stringent hydrolase selection criteria that based on protein length, the presence of three catalytic sites and predicted ORF completeness, the numbers of matches dropped to 19 from the entire database (Table 2.4). All sequences that fulfilled the criteria were derived from HMMsearch threshold value <1E-20, which indicated that a higher E-value threshold is likely retrieving a higher number of sequences that were either short, have incomplete catalytic sites or not a complete ORF. 13 of the sequences were obtained from the mixolimnion zone of Sanger-454 hybrid dataset, 3 from the 454 pyrosequencing assemblies and two matches from the Sanger reads (Table 2.4). Only one sequence was from the monimolimnion (169958875). Sequences from the 0.1 μm filter size were of utmost interest since the clone libraries harbouring the predicted genes are available. Information from the 454 pyrosequencing assemblies that constitutes data from 0.8 and 3.0 μm complement the data from 0.1 μm and revealed the diversity of microorganisms throughout the lake.

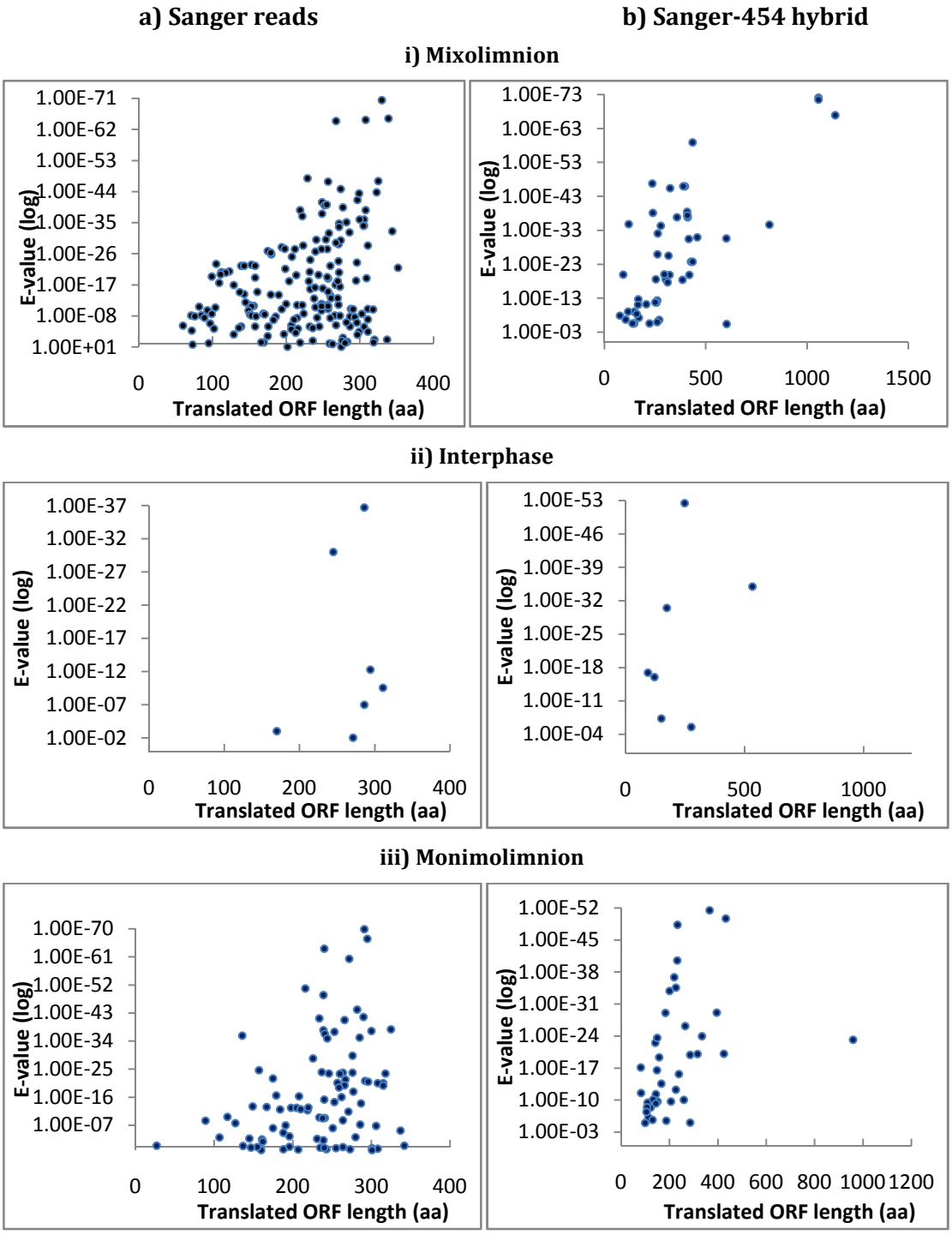


Figure 2.4: Distribution of the subtilase sequence matches in the 0.1 µm data. a) Sanger reads b) Sanger-454 hybrid in the three main zones of the lake according to the protein length and E-value. i) mixolimnion ii) interphase and iii) monimolimnion.

Table 2.4**:** List of the subtilase matches that passed the high stringency filtering criteria based on E-value, protein length, presence of catalytic sites and complete ORF prediction.

No	Sample ID	Sequence ID	Dataset ID	Filter size (µm)	Depth (m)
1	232	167813433	Sanger-454 hybrid	0.1	5
2	232	167774469	Sanger-454 hybrid	0.1	5
3	232	167861678	Sanger-454 hybrid	0.1	5
4	232	167890758	Sanger-454 hybrid	0.1	5
5	232	167865372	Sanger-454 hybrid	0.1	5
6	232	167703364	Sanger-454 hybrid	0.1	5
7	232	167718106	Sanger-454 hybrid	0.1	5
8	231	163343819	Sanger-454 hybrid	0.1	11.5
9	231	163430931	Sanger-454 hybrid	0.1	11.5
10	231	163154518	Sanger-454 hybrid	0.1	11.5
11	231	163539195	Sanger-454 hybrid	0.1	11.5
12	231	163120751	Sanger-454 hybrid	0.1	11.5
13	231	163128715	Sanger-454 hybrid	0.1	11.5
14	231	163497393	Sanger-454 hybrid	0.1	11.5
15	232	232_1113298458479	Sanger	0.1	5
16	232	232_1112308614602	Sanger	0.1	5
17	232	175873336	454 assemblies	0.8	5
18	228	169958875	454 assemblies	0.8	18
19	231	168645997	454 assemblies	3.0	11.5

2.3.2.1 Annotation of the Ace Lake subtilase sequences

All sequences derived from the HMMsearch were confirmed as serine protease, subtilisin or peptidase S8 through annotation in the KEGG, NCBI-nr and Swiss-Prot database except for 232_1113298458479, which was annotated as putative zinc metalloprotease in KEGG and as hypothetical protein in the NCBI-nr database.

2.3.2.2 Phylogenetic analysis of Ace Lake and commercially available subtilases

In order to determine the phylogenetic relationships amongst the conserved segments of the 19 subtilase protein sequences, a consensus Neighbor-Joining tree (Figure 2.5) was constructed. The bootstrap consensus tree was inferred from 1000 replicates (Felsenstein, 1985). The major clade that contained eight of the matches from the mixolimnion clustered together with uncharacterized serine proteases from marine *Actinobacteria* and plant pathogen *Actinobacteria* (*Clavibacter michiganensis* and *Leifsonia xyli*) respectively. The high bootstrap support (72-100%) suggested that the sequences were likely from *Actinobacteria* relative to *Clavibacter michiganensis* and *Leifsonia xyli*. Another sequence from the mixolimnion, 232_01|167774469, clustered with serine protease of the cyanobacterium, *Arthrospira* sp PCC 8005. 232_1113298458479 showed to be clustered together with low bootstrap value (40%) to *Oscillatoria* sp. PCC 6506. The only sequence from the anoxic monimolimnion, 228_08|169958875, was closely related to the anaerobic "*Candidatus* Cloacamonas acidaminovorans".

There were four sequences that were most closely related to *Gammaproteobacteria*-like proteases. 232_01|167890758 and 231_01|163128715 grouped together with extracellular protease of *Xanthomonas* sp. while 232_1112308614602 and 232_01|167703364 clustered with cold-adapted proteinase K-like protein of *Vibrio* sp PA44. 231_01|163343819 showed to be closely related to proteinase K-like protein but with a low bootstrap value. 231_01|163154518 and 232_01|167718106 made up a cluster with peptidase of a member of the *Chloroflexi* phylum, *Roseiflexus castenholzii* DSM 13941. Finally 231_30|168645997 formed the most distant group from the other Ace Lake sequences, clustering with subtilisin-like protease of *Polaribacter irgensii* 23-P (*Flavobacteria*). The phylogenetic tree also illustrated that none of the sequences from the Ace Lake metagenomic dataset, clustered together with the commercially available subtilase, subtilisin Savinase and subtilisin Calsberg, which were derived from *Bacillus* species.

2.3.2.3 Domain architecture analysis of the subtilase sequences

Domain architecture analysis was performed in CDD and CDart on the 13 unique subtilase sequences from the 'stringent group' and compared to the annotation in the MEROPS database to gain an insight into their physiological functions. Apart from the common domain for members of the subtilase superfamily which is peptidase_S8_S53 domain, some of the sequences had other specific domains either at their N-terminal or C-terminal

regions. The sequences that were most closely related to *Actinobacteria* (163430931, 163539195, 175873336, and 167865372) had domain architectures similar to T7 mycosin secretion system (T7SS-Mycosin) and were classified as belonging to the cytotoxin SubAB group in MEROPS. 163128715, which was annotated as proteins that originated from *Xanthomonas*, had two specific domains; subtilase uncharacterized subfamily 13 (peptidase_S8_13) and OmpA_family. 167703364, 232_1112308614602 and 163343819 were linked to diverse taxa but all sequences possessed the peptidase S8 PCSK9; a proteinase K-like proteins domain. 168645997, which was closely related to *Flavobacteria*, had the subtilase uncharacterized subfamily 9 (peptidase_S8_9) and Por secretion system C-terminal sorting domain (Por_secre_tail). The *Alphaproteobacteria* linked sequence; 232_1113298458479, contained a subtilase uncharacterized subfamily 4 (peptidase_S8_4) domain. Summary of the analysis were presented in the Table 2.5.

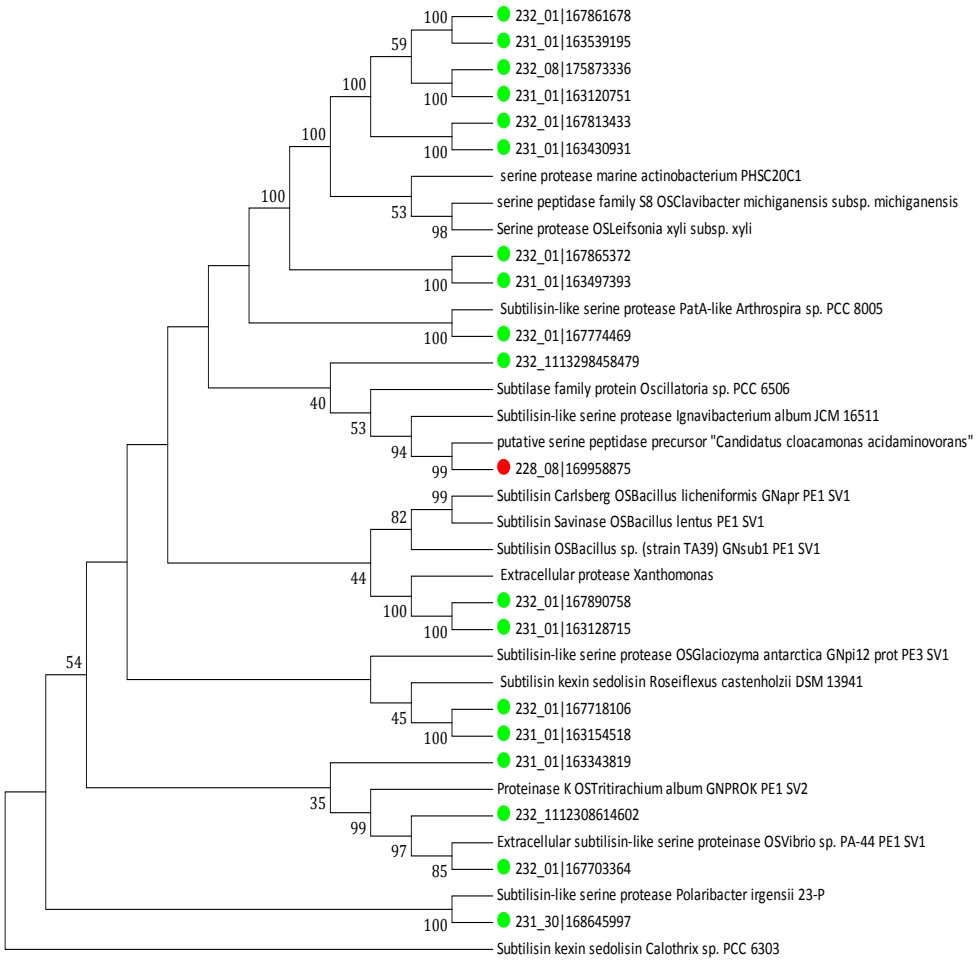


Figure 2.5: The phylogenetic tree of the subtilase-like protein sequences from Ace Lake as inferred using the Neighbor-Joining method. The bootstrap consensus tree was inferred from 1000 replicates. The green circles were sequences from the mixolimnion, the red circle was sequence from the monimolimnion. The details included in the sequence ID ie; 232_01, showed the sample depth ID and filter size respectively.

Table 2.5**:** Result of analysis of domain architecture for selected subtilase sequences.

No	ID	Closest Taxon	Domain architecture			MEROPS classification
			Specific domain	Superfamily	Multidomain	
1	231_01_163430931	*Actinobacteria*	Nil	peptidase_S8_S53	T7SS_mycosin	S08.121: CytotoxinSubAB
2	231_01_163539195	"	Nil	peptidase_S8_S53	T7SS_mycosin	S08.121: CytotoxinSubAB
3	232_08_175873336	"	peptidase_S8_Subtilisin_like	peptidase_S8_S53	T7SS_mycosin	S08.121: CytotoxinSubAB
4	232_01_167865372	"	Nil	peptidase_S8_S53	nil	S08.121: CytotoxinSubAB
5	231_01_163154518	*Chloroflexi*	peptidase_S8_S53	peptidase_S8_S53		S08.UPA: subfamily S8A unassigned peptidase
6	232_01_167774469	*Cyanobacteria*	Nil	peptidase_S8_S53	protease_PatA	S08.156: patA peptidase
7	231_01_163128715	*Xanthomonadaceae*	peptidase_S8_13-OmpA_family	peptidase_S8_S53-IPT-OmpA_like domain	nil	S08.UPA: subfamily S8A unassigned peptidase
8	232_01_167703364	*Deinococcus-Thermus*	peptidase_inhibitor_I9, peptidase_S8_PCSK9_Proteinase_K_like_proteins	peptidase_inhibitor_I9-peptidase_S8_S53	nil	S08.008 Mername-AA053 peptidase
9	232_1112308614602	*Deinococcus-Thermus*	peptidase_S8_PCSK9_Proteinase_K_like_proteins	peptidase_S8_S53	nil	S08.051 aqualysin 1
10	231_01_163343819	*Strongylocentrotus*	peptidase_S8_PCSK9_Proteinase_K_like_proteins	peptidase_S8_S53	nil	S08.A54 isp6 g.p
11	231_30_168645997	*Flavobacterium*	peptidase_S8_9,Por_secre_tail	peptidase_S8_S53, por_secre_tail	nil	S08.058 subtilisin-like peptidase
12	232_1113298458479	*Alphaproteobacteria*	Peptidase_S8_4	peptidase_S8_S53	nil	S08.UPA: subfamily S8A unassigned peptidase
13	228_08_169958875	Uncharacterized anaerobic bacterium	peptidase_S8_Subtilisin_like	peptidase_S8_S53		S08.026 nasp peptidase

2.3.3 Lipases sequences in the HMM search results

Lipolytic enzymes are very important and attractive research subjects because their multifunctional properties, such as broad substrate specificities and regiospecificities. The HMMsearch of four lipase groups (lipase 1, lipase 2, lipase 3 and lipase GDSL) from the Ace Lake metagenomic dataset has resulted with matches only to the two of the lipase groups; lipase 3 and lipase GDSL.

2.3.3.1 Selection of the lipase 3 sequences in the HMMsearch results

After the high stringency filtering criteria based on protein length, identification of the catalytic sites, predicted ORF completeness and unique sequence, the number of matches were reduced to only three from the entire database (Table 2.6). All sequences that fulfilled the criteria were derived from HMM threshold value <1e-20 except for 163182232 which was from E-value cut-off 1E-05.

Table 2.6: List of the lipase 3 sequences that passed the high stringency filtering criteria based on E-value, protein length, presence of catalytic sites and complete ORF prediction.

No	Sample ID	Feature ID	Dataset ID	Filter size(µm)	Depth(m)
1	232	167696458	Sanger-454 hybrid	0.1	5
2	231	163436800	Sanger-454 hybrid	0.1	11.5
3	231	163182232	Sanger-454 hybrid	0.1	11.5

2.3.3.1.1 *Annotation of the Ace Lake lipase 3 sequences*

All sequences from HMMsearch were annotated as lipase through comparison to the KEGG and NCBI-nr and Swiss-Prot databases except for 163182232, which was annotated as a hypothetical protein. The annotations for all three sequences indicated homology to eukaryote lipase. The closest homolog for 167696458 was to the lipase family protein of *Tetrahymena thermophila*, a unicellular organism, which is a genetic model for animals commonly found in fresh water environments (Eisen *et al.,* 2006). The closest homolog for 163436800 was to the tryglycerol lipase of choanoflagellate, *Monosiga brevicollis*.

2.3.3.2 Selection of the lipase GDSL sequences in the HMMsearch results

Analysis of the matches to lipase GDSL based protein length and E-value are shown in the scatter plot (Figure 2.6). The scatter plot of matches in the 0.1 μm of Sanger reads and Sanger-454 hybrid from the three zones of the Ace Lake indicated that the mixolimnion samples comprised of more sequences with protein length greater than 200 aa. Short reads were predicted as partial ORFs.

The number of matches decreased to 17 following the high stringency filtering process (Table 2.7). All sequences that fulfilled the criteria were derived from the 0.1 μm mixolimnion samples.

Table 2.7**:** List of the lipase GDSL sequences that passed the high stringency filtering criteria based on E-value, protein length, presence of catalytic sites and complete ORF prediction.

No	Sample ID	Sequence ID	Dataset ID	Filter Size(μm)	Depth (m)
1	232	167817518	Sanger-454 hybrid	0.1	5
2	232	167780571	Sanger-454 hybrid	0.1	5
3	232	167687840	Sanger-454 hybrid	0.1	5
4	232	167825930	Sanger-454 hybrid	0.1	5
5	232	167882604	Sanger-454 hybrid	0.1	5
6	232	167666764	Sanger-454 hybrid	0.1	5
7	232	232_1113297879091	Sanger	0.1	5
8	232	232_1113297996613	Sanger	0.1	5
9	232	232_1113316102348	Sanger	0.1	5
10	232	232_1113316024971	Sanger	0.1	5
11	231	163414796	Sanger-454 hybrid	0.1	11
12	231	163429503	Sanger-454 hybrid	0.1	11
13	231	163235684	Sanger-454 hybrid	0.1	11
14	231	163262734	Sanger-454 hybrid	0.1	11
15	231	231_1113297565557	Sanger	0.1	11
16	231	231_1113297793571	Sanger	0.1	11
17	231	231_1113289401511	Sanger	0.1	11

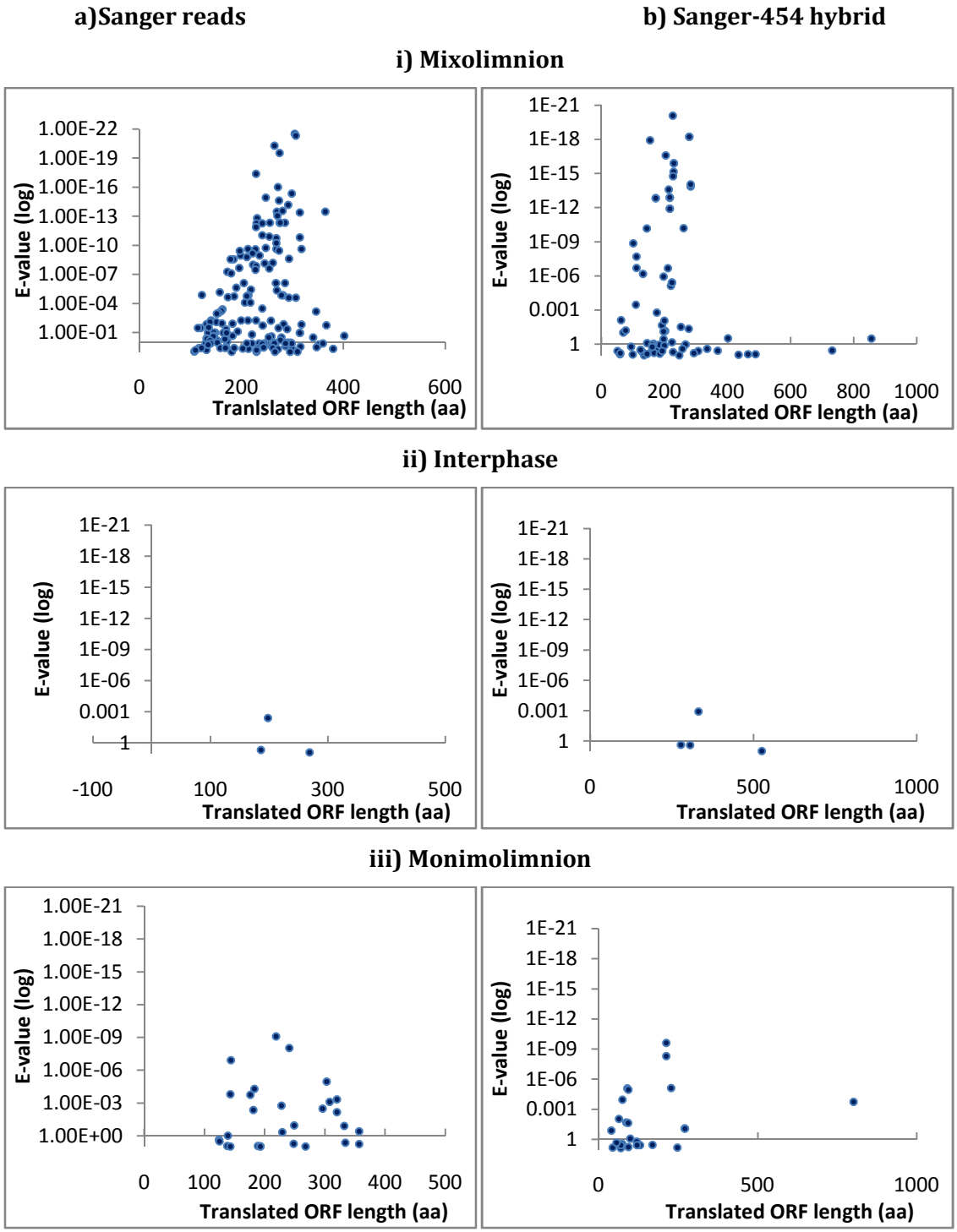


Figure 2.6: Distribution of lipase GDSL sequences matches in the 0.1 µm data. a) Sanger reads b) Sanger-454 hybrid in the three main zones of the lake according to the protein length and E-value. i) mixolimnion ii) interphase and iii) monimolimnion.

2.3.3.2.1 *Annotation of Ace Lake lipase GDSL sequences*

Annotation in KEGG, NCBI-nr and Swiss-Prot databases for each of the 17 selected matches from the HMM search were identified to confirm that the sequences were lipase GDSL. According to the annotation in the KEGG and NCBI-nr databases, 10 of the selected matches after high stringency filtering (Table 2.7) were lipase GDSL which were 167817518, 167687840, 167825930, 167882604, 232_1113316102348, 232_1113316024971, 163429503, 231_1113297565557, 231_1113289401511 and 231_1113297793571. Another four sequences, i.e., 167780571, 232_1113297879091, 163414796 and 163235684 were annotated as putative acyl coenzyme A thioesterase/aryl esterase. Two of the sequences, i.e., 167666764 and 232_1113297996613, were annotated as lysophospholipase. The remaining sequence, 163262734 was annotated as a hypothetical protein. Annotation by comparison to the Swiss-Prot database indicated that some of the matches in the search were not lipase but rather a glutamine-dependent NAD^+ synthetase (167817518), UDP-N-acetyl-enol-pyruvoyl glucosamine reductase (167882604) and DNA mismatch repair protein MutS (231_1113297565557 and 231_1113297793571).

2.3.3.2.2 *Phylogenetic analysis of lipase GDSL*

Phylogenetic relationships amongst the 17 predicted lipase GDSL sequences were analysed by constructing a consensus Neighbor-Joining tree. The sequences were generally divided into two major clades (Figure 2.7). The first clade was further divided into two subgroups. The first subgroup was represented by five Ace Lake sequences (231_1113297565557, 232|167825930, 232|167882604, 232_1113316102348, 232|167666764) that clustered with *Gammaproteobacteria* such as *Alcanivorax* sp. DG881 and *Pseudoalteromonas atlantica* T6c. Another subgroup was formed by 231_1113297793571 and lipase GDSL of *Kordia algicida* OT-1 (*Bacteroidetes*), while 231|163262734 and 232|167817518 formed another cluster with lipase GDSL of *Actinobacteria; Saccharopolyspora erythraea* NRRL 2338.

The second clade contained two sub-clusters with 100% bootstrap value. These were: (1) 232_1113297996613 and sialate O-acetylesterase of *Bacteroides fragilis* YCH46, (2) 232_1113316024971 and a lipolytic enzyme of *Bradyrhizobiaceae bacterium* SG-6C (*Alphaproteobacteria*). 232_1113297879091 and 231|163414796 have 100% bootstrap value with aryl esterase of "*Candidatus* Pelagibacter ubique" HTCC1002. There was another distant sequence in the same cluster, 231|163235684, which was closer to

arylesterase of *Desulfovibrio alaskensis* G20 (*Deltaproteobacteria*). Another lower confidence sub-cluster (57% bootstrap value) was apparent within the second clade that comprised of 232|167687840, 231_1113289401511, 231|163429503, lipase GDSL of *Francisella novicida* U112 (*Gammaproteobacteria*) and *Desulfobacula toluolica* Tol2 (*Deltaproteobacteria*). Finally, 232|167780571 formed the most distant related clade with arylesterase of *Rhodobacterales bacterium* HTCC2255 and pancreatic lipase.

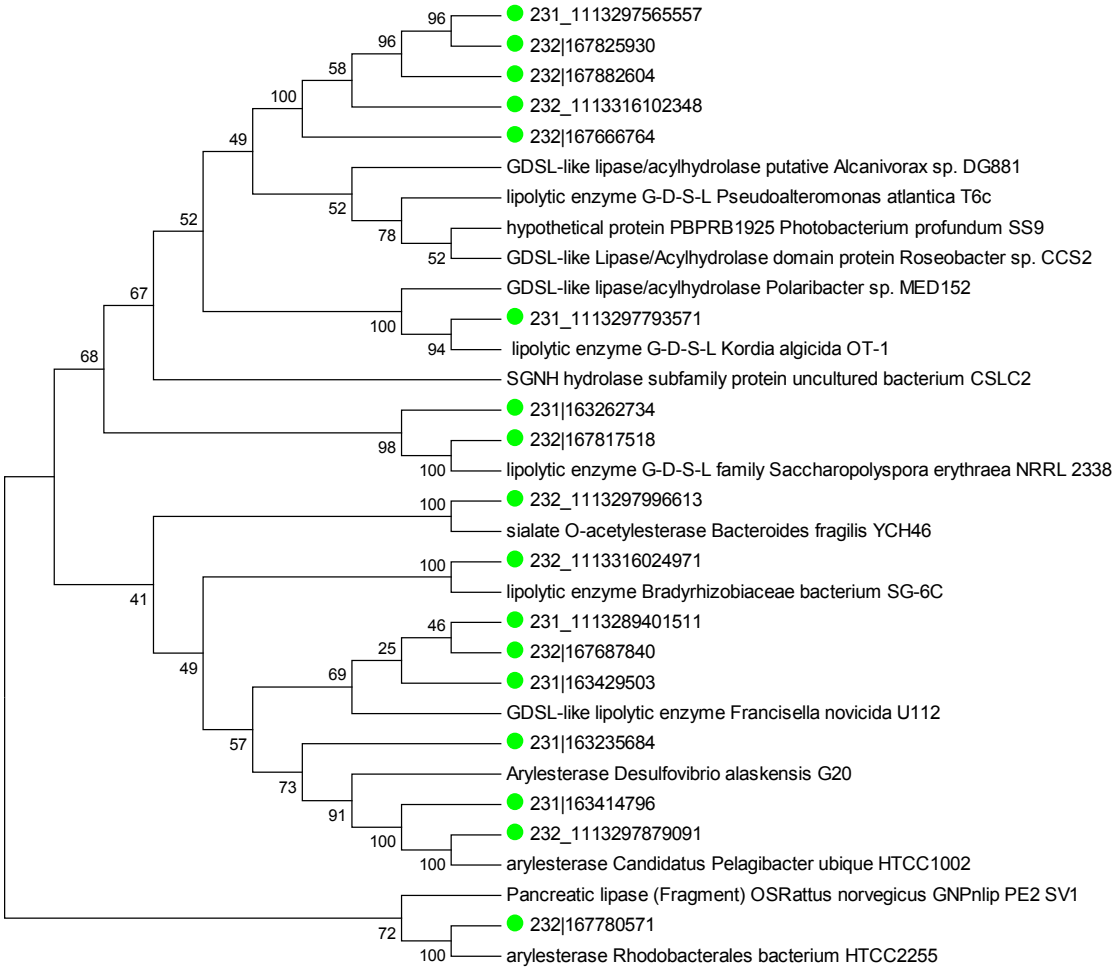


Figure 2.7**:** Phylogenetic tree of the Ace Lake metagenome-derived lipase GDSL sequences. The green circles were sequenced from the oxic mixolimnion. The details included in the sequence ID i.e., lcl|232_01, showed the sample depth ID.

2.3.3.2.3 *Conserved domain architecture analysis*

Domain architecture analysis was performed on the 17 selected sequences from the 'stringent group' (Table 2.7) to search for the presence of novel domain. All sequences had the SGNH hydrolase superfamily domain, a common domain for lipase GDSL sequences. Members of the SGNH hydrolase 1 superfamily domain were found in the sequences related to *Actinobacteria* and *Alphaproteobacteria* SGNH_hydrolase subfamily. FeeA_FeeB-

like domain was detected in at least four sequences related to *Gammaproteobacteria* (167825930, 167882604, 167666764, 231_1113297565557), and two sequences that were related to *Flavobacteria* and *Betaproteobacteria* (231_1113297793571 and 232_1113316102348 respectively). The FeeA and FeeB genes are part of a biosynthetic gene cluster and may participate in the biosynthesis of long chain N-acyltyrosines by providing saturated and unsaturated fatty acids, which in turn are loaded onto the acyl carrier protein FeeL (Brady *et al.,* 2002). Finally, the two sequences associated with *Alphaproteobacteria* (232_1113297879091, 163414796) and another sequence associated with *Deltaproteobacteria (*163235684) have a lysophospholipase L1-like domain, a subgroup of SGNH-hydrolases domain.

Table 2.8**:** Result of analysis of domain architecture for selected lipase GDSL sequences.

ID	Closest related taxon	Specific domain	Superfamily
167817518	*Actinobacteria*	SGNH_hydrolase_like_1	SGNH_hydrolase superfamily
163262734	*Actinobacteria*	SGNH_hydrolase_like_1	SGNH_hydrolase superfamily
231_1113297793571	*Flavobacteria*	FeeA_FeeB_like	SGNH_hydrolase superfamily
167825930	*Gammaproteobacteria*	FeeA_FeeB_like	SGNH_hydrolase superfamily
167882604	*Gammaproteobacteria*	FeeA_FeeB_like	SGNH_hydrolase superfamily
167666764	*Gammaproteobacteria*	FeeA_FeeB_like	SGNH_hydrolase superfamily
231_1113297565557	*Gammaproteobacteria*	FeeA_FeeB_like	SGNH_hydrolase superfamily
232_1113316102348	*Betaproteobacteria*	FeeA_FeeB_like	SGNH_hydrolase superfamily
167780571	*Deltaproteobacteria*	-	SGNH_hydrolase superfamily
167687840	*Gammaproteobacteria*	-	SGNH_hydrolase superfamily
231_1113289401511	*Deltaproteobacteria*	-	SGNH_hydrolase superfamily
163429503	*Gammaproteobacteria*	-	SGNH_hydrolase superfamily
232_1113297996613	*Sphingobacteria*	-	SGNH_hydrolase superfamily
232_1113316024971	*Alphaproteobacteria*	SGNH_hydrolase	SGNH_hydrolase superfamily
163235684	*Deltaproteobacteria*	Lysophospholipase_L1_like	SGNH_hydrolase superfamily
232_1113297879091	*Alphaproteobacteria*	Lysophospholipase_L1_like	SGNH_hydrolase superfamily
163414796	*Alphaproteobacteria*	Lysophospholipase_L1_like	SGNH_hydrolase superfamily

2.3.4 Selection of the GH13 sequences in the HMMsearch results

The high number of GH13 matches sequences were further analysed based on their length and E-value. Scatter plots of the number of matches in the 0.1 μm sample of Sanger-454 hybrid and Sanger reads from the three main zones of the Ace Lake (mixolimnion (5 and 11 m), oxic-anoxic interphase (12.7 m) and monimolimnion (14, 18, 23 m)) are shown (Figure 2.8).In comparison to the matches in the Sanger reads, the plots indicated some of the GH13 matches in the Sanger-454 pyrosequencing hybrid consisted of longer translated ORFS. This was apparent in the mixolimnion and interphase zones, where the longest translated ORF have more than 1000 aa. In contrast, most of the proteins predicted from Sanger reads were shorter (less than 300 aa). Results from analysis with MetagGene indicated that short reads were predicted as partial ORFs.

After applying the high stringency filtering process, 27 GH13 sequences were obtained from the metagenome dataset (Table 2.9). All sequences were from the mixolimnion and interphase zones and had E-value < 1E-20, except for 167853847, 167817168, 167817172 and 163239763 that had E-value < 1E-5.

Table 2.9: List of the GH13 sequences that passed the higher stringency filtering criteria based on E-value, protein length, presence of catalytic sites and complete ORF prediction.

No.	Sample ID	Sequence ID	Dataset ID	Filter size(µm)	Sample depth (µm)
1	232	167733578	Sanger-454 hybrid	0.1	5.0
2	232	167822374	Sanger-454 hybrid	0.1	5.0
3	232	167733580	Sanger-454 hybrid	0.1	5..0
4	232	167687568	Sanger-454 hybrid	0.1	5.0
5	232	167667080	Sanger-454 hybrid	0.1	5.0
6	232	167715284	Sanger-454 hybrid	0.1	5.0
7	232	167853847	Sanger-454 hybrid	0.1	5.0
8	232	167817168	Sanger-454 hybrid	0.1	5.0
9	232	167817172	Sanger-454 hybrid	0.1	5.0
10	231	163470518	Sanger-454 hybrid	0.1	11.5
11	231	163360988	Sanger-454 hybrid	0.1	11.5
12	231	163470516	Sanger-454 hybrid	0.1	11.5
13	231	163427827	Sanger-454 hybrid	0.1	11.5
14	231	163488553	Sanger-454 hybrid	0.1	11.5
15	231	163277673	Sanger-454 hybrid	0.1	11.5
16	231	163277667	Sanger-454 hybrid	0.1	11.5
17	231	163239763	Sanger-454 hybrid	0.1	11.5
18	232	175711712	454 assemblies	0.8	5.0
19	232	176222994	454 assemblies	0.8	5.0
20	232	175711710	454 assemblies	0.8	5.0
21	232	175839179	454 assemblies	3.0	5.0
22	231	167541905	454 assemblies	0.8	11.5
23	231	167562621	454 assemblies	0.8	11.5
24	231	167588808	454 assemblies	0.8	11.5
25	231	168627242	454 assemblies	3.0	11.5
26	230	167494038	454 assemblies	0.8	12.7
27	230	178055561	454 assemblies	3.0	12.7

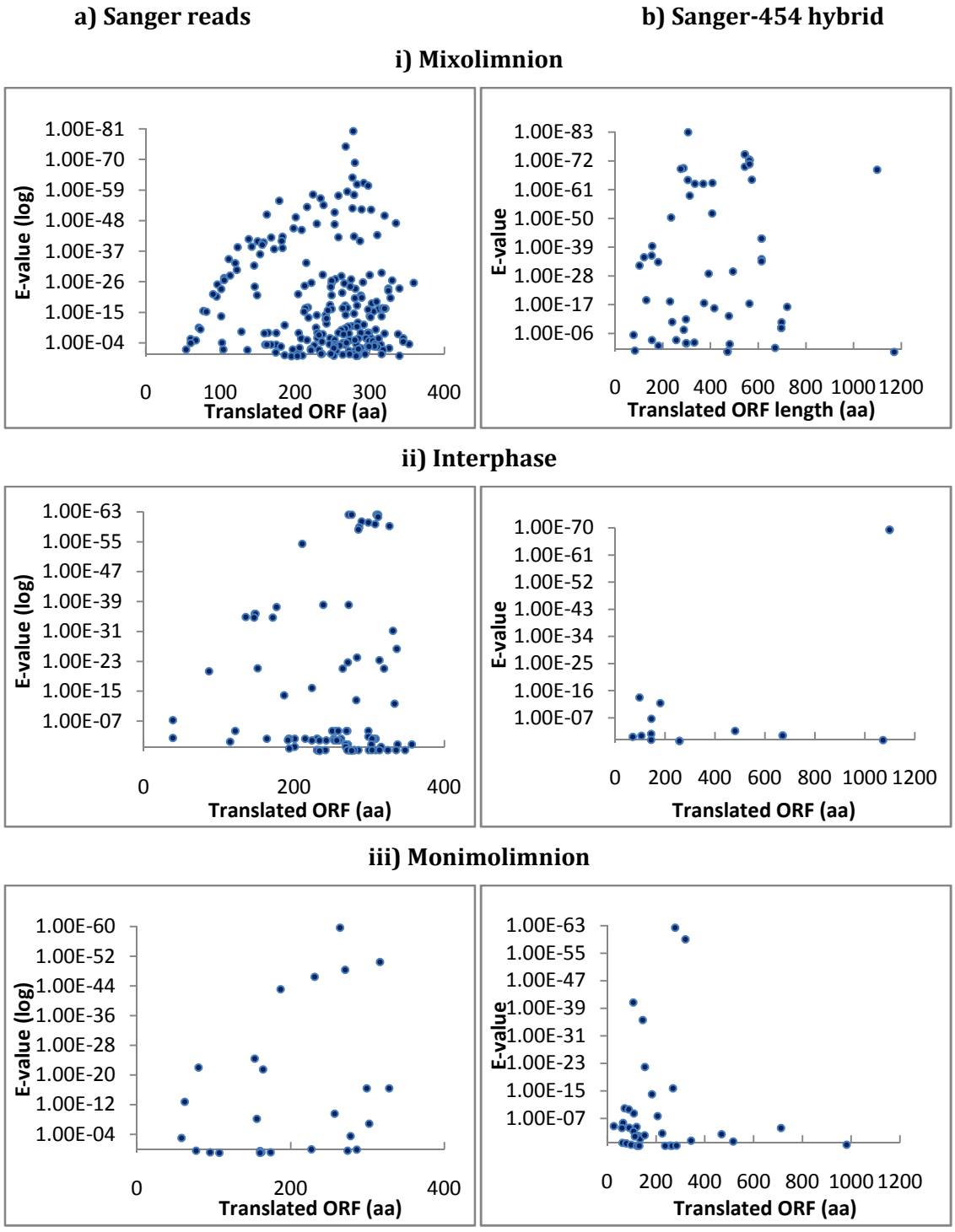


Figure 2.8: Distribution of GH13 sequence matches in the 0.1 µm data**:** a) Sanger reads b) Sanger-454 hybrid in the three main zones of the lake according to the protein length and E-value. i) mixolimnion ii) interphase and iii) monimolimnion.

2.3.4.1 Annotation of the Ace Lake GH13 sequences

Annotation in KEGG, NCBI-nr and Swiss-Prot databases for each of the 27 selected matches from the HMMsearch were identified to confirm that the sequences were GH13 sequences. Apart from that, the annotation differentiated whether the GH13 sequences were α-glucosidase, α-amylase, trehalose synthase or glycogen branching/debranching enzyme. According to the annotation in the KEGG and NCBI-nr databases, 11 of the sequences were annotated as α-glucosidase (167733578, 167822374, 167733580, 167687568, 167667080, 175711712, 163470518, 163360988, 163470516, 163427827 and 167562621). Another five sequences were annotated as α-amylase (176222994, 175839179, 167541905, 167588808 and 168627242). There were four sequences that were annotated as trehalose synthase (167715284, 163488553, 167494038 and 178055561). Two sequences from the interphase zone, i.e., 167494038 and 178055561, were annotated as trehalose synthase related to the green sulfur bacterium, *Chlorobium phaeovibrioides.* There were five sequences that were annotated as glycogen branching/ debranching enzymes which were 167817168, 167817172, 163239763, 163277667 and 163277673. The remaining two sequences, i.e., 167853847 and 175711710 were annotated as a conserved hypothetical lipoprotein and a hypothetical protein, respectively.

2.3.4.2 Plylogenetic analysis of Ace Lake GH13 sequences

Phylogenetic analysis of GH13 proteins showed the sequences were divided according to the five members of the α-amylase family, namely α-glucosidase, α-amylase, trehalose synthase, amylosucrase and glycogen branching/debranching enzyme (Figure2.9). 232_01|167733578, 231_01|163470518, 232_01|167733580, 231_01|163470516 and 232_01|167667080 clustered with α-glucosidase of *Clavibacter michiganensis* subsp. *michiganensis* NCPPB 382 (*Actinobacteria*) and marine actinobacterium PHSC20C1 and *Thermobifidafusca* YX (*Actinobacteria*). 232_01|167822374, 231_01|163360988, 232_01|167687568 and 231_01|163427827 formed a separate cluster with no known homologous proteins. A cluster of sequences (232_08|175711710, 231_08|167562621, and 232_08|175711712) were related to oligo-1,6-glucosidase of *Cyanothece* sp. CCY0110, (*Cyanobacteria*), all of which originated from 0.8 μm mixolimnion. The high bootstrap value with cyanobacterial sequences indicated their similarity to Ace Lake *Cyanobacteria* that commonly found in the surface.

There were four Ace Lake GH13 sequences in the trehalose synthase cluster. 231_01|163488553 clustered with 100% bootstrap value to the trehalose synthase of an actinobacterium, *Leifsoniaxyli* CTCB07. 232_01|167715284 clustered with 100% bootstrap value to thetrehalose synthase of a green sulfur bacterium, *Chlorobium phaeovibrioides* DSM 265. Another two sequences in the cluster, 230_08|167494038, and 230_30|178055561 which were derived from the 0.8 and 3.0 μm filter of the interphase zone have lower bootstrap value (34%) indicating their distant relationship to *Chlorobium phaeovibrioides* DSM 265.

In a different cluster, 231_08|167541905, and 231_30|168627242 grouped with amylosucrase of *Cyanobacteria*; *Synechococcus* sp. WH 5701 and a sulfur-oxidizing bacterium isolated from deep sea hydrothermal vents of class *Gammaproteobacteria*; *Thiomicrospiracrunogena XCL*-2. Two predicted sequences of α-amylase 231_08|167588808 and 230_08|176222994, grouped with α-amylase of a *Cyanobacteria*; *Acaryoschloris marina*. Another two clusters of α-amylase comprised of 232_30|175839179 that grouped with 100% bootstrap value to α-amylase of *Verrucomicrobiae* bacterium DG1235 and 232_01|167853847 which grouped together with α-amylase of a member of the *Bacteroidetes*, *Flavobacteria bacterium* BBFL7.

The remaining sequences formed the clusters of glycogen branching/debranching enzymes. 232_01|167817168 and 231_01|163277673 clustered with glycogen branching enzyme of *Arthrobacteraurescens* TC1 (*Actinobacteria*). Finally, 231_01|163277667, 232_01|167817172 and 232_01|163239763 grouped with glycogen debranching enzyme of *Actinobacteria*, *Mycobacterium vanbaalenii* PYR-1 and *Acidothermus cellulolyticus* 11B.

2.3.4.3 Domain architecture analysis of Ace Lake GH13 sequences

Domain architecture analysis was performed on the 18 selected GH13 sequences from the 'stringent group' (Table 2.9) as one of the search for functional novelty indicator. Apart of the common catalytic domain for the α-amylase family which is AmyAc superfamily, each of the sequence has specific domain from which their function can be inferred. Four types of specific domains were found in the sequences linked to *Actinobacteria*: the α-amylase catalytic domain that was found in the oligo-1,6-glucosidase and trehalose synthase (AmyAc_OligoGlu_TS); α-amylase catalytic domain that was found in the trehalose synthase (AmyAc_TreS); the catalytic domains of the glycogen branching enzyme (AmyAc_Glg_BE) and the catalytic domain of the glycogen debranching enzymes

(AmyAc_Glg_debranch). Both glycogen branching (GBE) and debranching enzymes (GDE) have specific domain at their N-terminal that was known as early set domain (E-set). The difference between branching and debranching enzymes was the presence of multidomain of glycogen branching enzyme (PRK05402) in the former and the presence of the bacterial isoamylase domain (isoamylase_N) and module of multidomain glycogen debranching enzyme (GlgX-debranch) in the latter.

AmyAc_OligoGlu_TreS and AmyAc_TreS domains were also detected in the sequences related to *Chlorobium* while AmyAc_OligoGlu domain was detected in the sequences related to *Cyanobacteria*. Other specific domains found in the sequences related to *Cyanobacteria* were the α-amylase catalytic domain, found in the bacterial and fungal α-amylases (AmyAc_bac_fung_AmyA), and the α-amylase catalytic domain found in amylosucrase (AmyAc_Amylsucrase).

The α-amylase catalytic domain belonging to an uncharacterized protein family (AmyAc_4) was detected in sequence associated with *Bacteroidetes*. Finally, α-amylase catalytic domain found in archaeal and bacterial α-amylases (AmyAc_arch_bac_AmyA) were detected in the sequence related to *Verrumicrobiae*. All of the detected domains confirmed the annotation of each sequence described in the previous sections (Table 2.10).

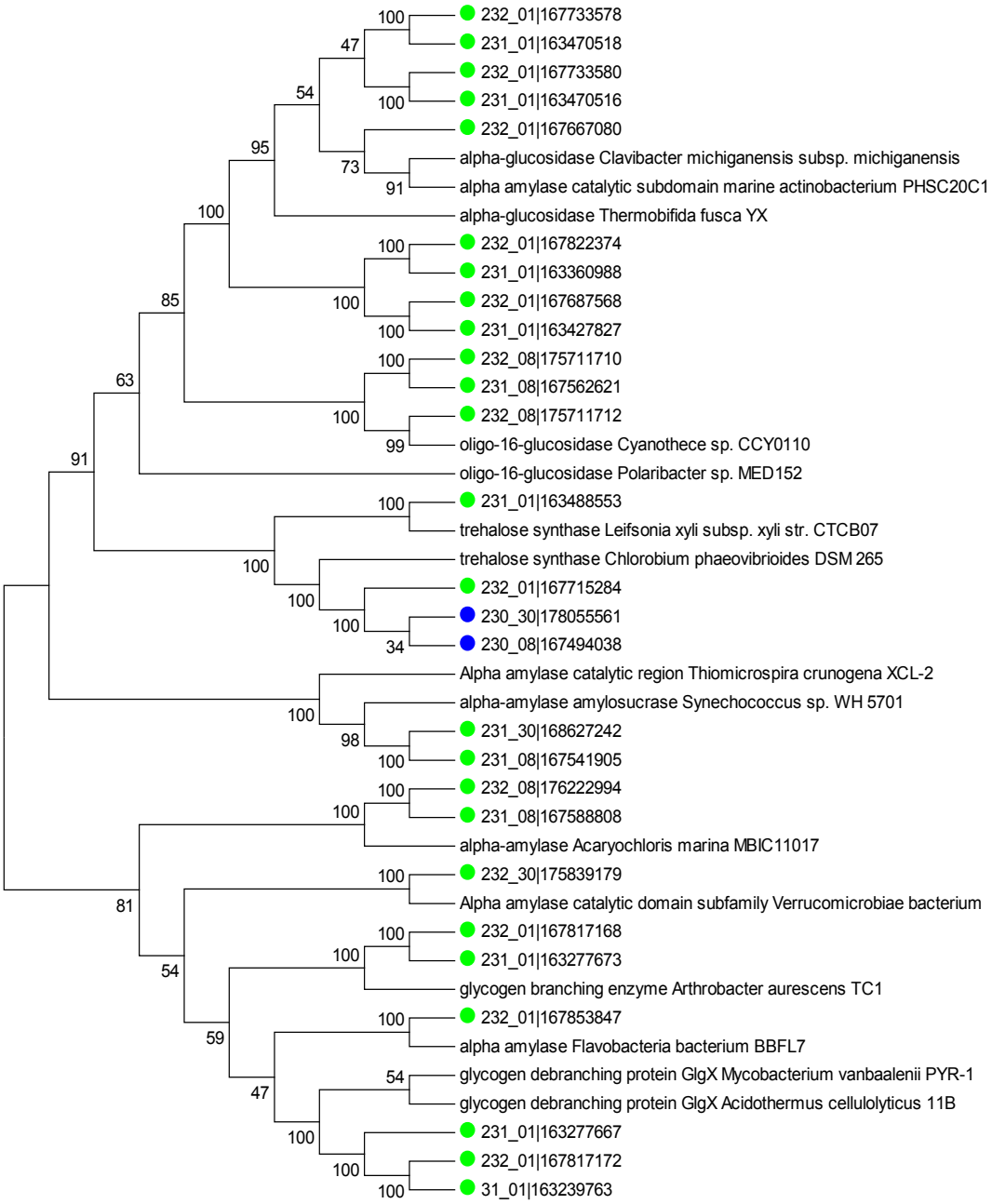


Figure 2.9: Phylogenetic tree of GH13 sequences derived from the Ace Lake metagenome. The phylogenetic tree was inferred using the Neigbor-Joining method. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches. The green circles were sequences from the oxic mixolimnion and the blue circles were sequences from the interphase sample. The details included in the sequence ID i.e., 232_01 showed the sample depth ID and the filter size respectively.

Table 2.10**:** Analysis of domain architectureof the GH13 sequences.

ID	Closest related Taxon	Spesific domain	Superfamily	Multidomain
167733578	*Actinobacteria*	AmyAc_OligoGlu_TS	AmyAc	trehalose_treC
167667080	*Actinobacteria*	AmyAc_OligoGlu_TS	AmyAc	trehalose_treC
167822374	*Actinobacteria*	AmyAc_OligoGlu_TS	AmyAc	nil
167733580	*Actinobacteria*	AmyAc_OligoGlu_TS	AmyAc	nil
167687568	*Gammaproteobacteria*	AmyAc_OligoGlu_TS	DUF_3459	trehalose_treC
167715284	*Chlorobium*	AmyAc_OligoGlu_TS	Glycosyltransferases, GH31, BLE	treS-Nterm_treS-Cterm
163488553	*Actinobacteria*	AmyAc_TreS	AmyAC_family	treS_nterm
167494038	*Chlorobium*	AmyAc_TreS	AmyAc_family superfamily	treS_nterm
178055561	*Chlorobium*	AmyAc_TreS	AmyAc_family superfamily	treS_nterm
167817168	*Actinobacteria*	E_set_GBE_prok_N, AmyAc_Glg_BE,*A*_amylase_C	E_set, AmyAC_family	PRK05402- glycogen branching enzyme
163277667	*Actinobacteria*	E_set_GDE_Isoamylase_N, AmyAc_Glg_debranch	E_set, AmyAC_family	glgX_debranch
163239763	*Actinobacteria*	E_set_GDE_Isoamylase_N, AmyAc_Glg_debranch	E_set, AmyAC_family	glgX_debranch,
168627242	*Cyanobacteria*	AmyAc_Amylsucrase	AmyAc_family superfamily	treS_nterm
176222994	*Cyanobacteria*	AmyAc_bac_fung_AmyA	AmyAc_family superfamily	PRK09441-cytoplasmic alphaamylase
175711710	*Cyanobacteria*	AmyAc_OligoGlu	AmyAc_family superfamily	AmyA
175711712	*Cyanobacteria*	AmyAc_OligoGlu	AmyAc_family superfamily	trehalose_treC
175839179	*Verrucomicrobiae*	AmyAc_arch_bac_AmyA	AmyAc_family superfamily	AmyA
167853847	*Bacteroidetes*	AmyAc_4	AmyAc_family superfamily	AmyA

2.4 Discussion

2.4.1 Implication of DNA sequencing and processing techniques towards the number of matches in the HMMsearch

For sample of 0.1 μm, in respect to the selection criteria, more matches of subtilase, lipase and glycosyl hydrolase were found in the assembled data of mixolimnion compared to the interphase and deep monimolimnion. This was in line with the decrease of annotation percentage for predicted genes with depth in the Ace Lake metagenome dataset, dropping to <3% in the deepest zone at 23 m (Lauro *et al.,* 2011). In comparison of assembled and unassembled data, more matches were detected in the former than the latter. This result might be implicated by the difference of sequencing methods and data processing applied to generate each dataset. Sequences in the 0.1 μm of unassembled data were generated by Sanger sequencing while sequences in the 0.1 μm of assembled data were from combination of Sanger and 454 pyrosequencing technologies. Even though the HMMsearch in the unassembled data of 0.1 μm primarily showed higher number of matches, they were removed in the selection process because they were comprised of less than 200 aa (or 300 aa for glycosyl hydrolase), were partial ORFs, or had incomplete catalytic sites. It was also worthwhile to mention that the abundance of the matches in both the assembled and unassembled data cannot be directly compared since the assembly process combined sequence reads into contiguous stretches of DNA called contigs, based on sequence similarity between reads (Kunin *et al.,* 2008) and contributed to the reduction of number of translated ORFs in the assembled datasets.

Lower numbers of matches were detected in the 0.8 and 3.0 μm samples which comprised of only 20.6% of the total complete ORFs identified. Sequence in the assembled data of 0.8 and 3.0 μm were generated by 454 pyrosequencing technology. While the advantages of 454 pyrosequencing over Sanger sequencing were its lower cost and no requirement for cloning, its main disadvantage has been its overall read length: ~400bp (Uchiyama and Miyazaki, 2010). Reads of this length present additional challenges for assembly and gene calling (Kunin *et al.,* 2008) and posed similar problems for detecting complete ORFs for hydrolase in this project. Therefore, in this project, it showed that the combination of Sanger and pyrosequencing had an advantage for the purpose of

identification of subtilase, lipase and GH13 complete ORFs. This approach produced longer contigs and increased the probability to obtain complete ORFs. A similar approach (combination of Sanger and 454 pyrosequencing) has been determined by Goldberg *et al.* (2006) to be able to produce high quality cost-effective assemblies of small marine microbial genomes.

2.4.2 The matches of hydrolases were linked to the dominant taxa in the Ace Lake environment

The degradation and turnover of various materials are a continuous process mediated by the action of a variety of microorganisms. Consistent with higher representatives of HMMsearch results from the upper-oxic zones (mixolimnion), the detected subtilase, lipase GDSL and α-amylase family enzymes matches were mainly linked to *Actinobacteria*, *Proteobacteria*, *Cyanobacteria* and *Bacteroidetes*. According to Lauro *et al.* (2011), in the mixolimnion zones of the Ace Lake, the bacterial members responsible for remineralisation of particulate organic carbon to DOC include members of the *Flavobacteria* and *Gammaproteobacteria*, which are likely to be particle-associated copiotrophs. Heterotrophic conversion would be further performed by free-living, oligotrophic *Actinobacteria* and members of the SAR11 clade (Lauro *et al.,* 2011). Annotation of the identified hydrolase sequences from mixolimnion samples as related to *Actinobacteria* was consistent with over representation of *Actinobacteria* that was associated with Luna cluster in the upper-oxic zones of the Ace Lake (Lauro *et al.,* 2011). This cluster was mainly represented in freshwater ultramicrobacteria (Hahn *et al.,* 2003) and played a greater role in the bacterial assemblages of the oxygenated water layers than in the interphase and anoxic aquatic environments (Salcher *et al.,* 2010, Lauro *et al.,* 2011). The study of spatiotemporal distribution and activity patterns of bacteria from three phylogenetic groups in an oligomesotrophic alpine lake also indicated that *Actinobacteria* was the most active amino acid incorporating bacteria regardless of their small size (Salcher *et al.,* 2010).

Nevertheless, as expected, α-amylase family enzymes sequences derived from the interphase zone were linked to the green sulfur bacterium of the genus *Chlorobium*, that was found dominated that particular Ace Lake's zone (Ng *et al.,* 2010, Lauro *et al.,* 2011). Other α-

amylase family enzymes sequences identified in the 0.8 and 3.0 μm samples were associated with *Cyanobacteria,* which dominated the 0.8 and to a lesser extent, the 3.0 μm samples of the upper-oxic zones of the lake (Lauro *et al.,* 2011). *Cyanobacteria* of the genus *Synechococcus* represent a unique and highly adapted clade in the stable water columns of some saline Antarctic lakes such as Ace Lake (up to 8 x 10^6 cell mL^{-1}), Lake Abraxas (1.5 x 10^7 cells mL^{-1}) and Pendant Lake (max. 1.5 x 10^7 cells mL^{-1})(Powell *et al.,* 2005).

The only sequence from the anoxic zones (169958875), identified as subtilase, was linked to WWE_1 candidate division that deeply branching from the phylum *Spirochaetes* and was provisionally classified as "*Candidatus* Cloacamonas acidaminovorans"(Pelletier *et al.,* 2008). *In silico* proteome analysis indicated that this bacterium might derive most of its carbon and energy from the fermentation of amino acids (Pelletier *et al.,* 2008).

2.4.3 Conserved domain architecture of subtilase inferred probable biological function of the enzymes

Amongst the three enzymes sequences, the domain architecture analysis of the subtilase resulted in interesting findings in terms of the enzyme function. In bacteria, subtilisin functions in diverse processes such as in cellular nutrition uptake and host invasion (Cheng *et al.,* 2002), facilitating the maturation of diverse polypeptides (Siezen *et al.,* 2007), bacteriocins like lantibiotics (Siezen *et al.,* 1996), and extracellular adhesions and germination of spores (Shimamoto *et al.,* 2001). Most of the known subtilisins are multidomain polypeptides and consists of a protease domain accompanied by one or more co-existing domains (Geer *et al.,* 2002) which also counts for the diversity of their functions (Tripathi and Sowdhamini, 2008).

Analysis in CDD (Table 2.5) indicated the presence of a T7SS-Mycosin multidomain module in only those subtilase sequences that were associated with *Actinobacteria*. T7SS-mycosin is a secretion system that functions in transporting *Mycobacterium tuberculosis* protein virulence factors that cause tuberculosis in human (Stanley *et al.,* 2003, Abdallah *et al.,* 2007). 163539195, 163430931 and 175873336 have similar domain architecture with *MycP1*, a subtilisin-like protease identified in the T7SS, which involved in the early replication and full virulence of the pathogen in host cell (Ohol *et al.,* 2010). There was an

increasing number of observations indicated that planktonic *Actinobacteria* indigenous to freshwater systems are less vulnerable to protistan predation and viral lysis than other taxa of freshwater bacterioplankton (Horňák *et al.,* 2005, Tarao *et al.,* 2009). So far, the Gram-positive membrane cell structures were known to be responsible for protection (Horňák *et al.,* 2005, Tarao *et al.,* 2009). While the abundance of *Actinobacteria* in the mixolimnion of the Ace Lake has been linked to the protein remineralisation role (Lauro *et al.,* 2011), the similarity of 163539195, 163430931 and 175873336 domain architecture with *MycP1* suggesting there was a possibility that the indigenous *Actinobacteria* species in Ace Lake have specific virulence factor involving subtilase that function as a protective measure against predation. However, there was no experimental data for multifunction subtilase linked to the *Actinobacteria* of Luna cluster to support this statement. Nevertheless, there was a report of a multifunctional subtilase, *Streptococcal* SCP, which has a specific role in the inactivation of the human phagocyte chemotaxin C5, as well as playing an important role as an invasin that assists in epithelial cell invasion (Cheng *et al.,* 2002).

Another subtilase sequence that has interesting domain architecture is 163128715. Even though had high sequence identity to *Xanthomonas campesteris* (Liu *et al.,* 1990) and other uncharacterized serine protease of *Xanthomonadaceae* species, 163128715, has a peptidoglycan binding domains similar to the C-terminal domain of outer membrane protein (OmpA_C_like) instead of a common pre-peptidase C-terminal domain (PPC) in its C-terminal extension. In most subtilases of *Xanthomonas,* this PPC domain is normally found at the C-terminal of secreted bacterial peptidases and does not present in the active peptidase due to C-terminal excision during the maturation process. Further search for the similar domain architecture of other serine protease of *Xanthomonas* and *Stenotrophomonas* species resulted in only two groups that showed almost similar domain architecture thatwas lacking a PPC domain. The first group was a serine protease (Lee *et al.,* 2005a), and consisted of an autotransporter-associated beta strand repeat domain (cl15373) in between peptidase_S8_S53 and outer membrane (OM) channel domains. Another group, serotype 1 specific antigen has both peptidase_S8_S53 and OM channel domains. Both autotransporters and OM_channel domain were known as translocator domain that mediate the secretion of virulence-related protein of Gram-negative bacteria (Oomen *et al.,* 2004).

The cell envelope of Gram-negative bacteria is composed of two membranes, the inner and the outer membrane, which are separated by the peptidoglycan-containing periplasm. A number of proteins interact noncovalently with the cell wall through a periplasmic peptidoglycan binding domain that is widespread in Gram-negative bacteria (Parsons *et al.,* 2006). It was found that 163128715 OmpA_C-like domain was homologous to the C-terminal sequences of the reduction-modifiable protein M (RmpM) in *Neisseria meningitides* (Grizot and Buchanan, 2004), outer membrane protein (Rv0899) from *Mycobacterium tuberculosis* (Li *et al.,* 2012) and peptidoglycan-associated lipoprotein (Pal) of *Haemophilus influenza* (Parsons *et al.,* 2006) that were linked to their role as a peptidoglycan binding domain. Since 163128715 lacked an autotransporter domain, any role in pathogenesis can likely be excluded. As it lacks a signal peptide and has a peptidoglycan binding domain, Subt8715 most likely was an intracellular membrane protein that functions in the maturation of other proteins or peptides.

In line with the role of *Flavobacteria* in remineralisation of particulate organic carbon to DOC in Ace Lake, a sequence associated with *Flavobacteria* (168645997) has a Por_secre_tail domain at its C-terminal. This domain was linked to its function of protein sorting to the outer membrane. Involved in the Por secretion system (PorSS), it was first identified in the two members of the *Bacteroidetes* phylum, the gliding bacterium *Flavobacterium johnsoniae* and the nonmotile oral pathogen *Porphyromonas gingivalis* and has different components compared to those of other well studied bacterial secretion systems (McBride and Zhu, 2013). In *P. gingivalis* this secretion system was responsible for secretion of major extracellular cysteine proteinases, Arg-gingipains (Rgps) and Lys-gingipain (Shoji *et al.,* 2011). Further comparative analysis of 37 genomes of members of the phylum *Bacteroidetes* revealed the widespread occurrence of the PorSS genes (McBride and Zhu, 2013) suggesting similar secretion system were being used by indigenous *Bacteroidetes* members in Ace Lake for secretion of extracellular subtilase.

There were three sequences that had specific proteinase K-like subtilase domain (232_1112308614602, 232_1112308614602 and 163343819). The study of subtilase of this group is quite comprehensive especially in term of temperature adaptation (Arnórsdóttir *et al.,* 2005, Sakaguchi *et al.,* 2007), as mesophilic, thermophilic, and psychrophilic

representatives have been characterized. The physiological function of this group was not clear. However, the known proteinase K-like subtilase has broad substrate specificity.

2.4.4 Functional potential of lipase GDSL in the Ace Lake aquatic environment

From analysis of the conserved domain and comparison to the annotation through NCBI-nr, Swiss-Prot and KEGG database, five lipase GDSL sequence annotated as GDSL lipolytic enzymes and one sequence annotated as lysophospholipase LI and related esterase-like protein have FeeA and FeeB-like specific domain (Table 2.8). All of the sequences were associated with Gram-negative *Gammaproteobacteria*, *Deltaproteobacteria* and *Bacteroidetes*. These FeeA and FeeB-like proteins were part of a biosynthetic gene cluster and may participate in the biosynthesis of a long chain N-acyl amino acids by providing saturated and unsaturated fatty acids (Brady *et al.,* 2002). The long chain of N-acyl amino acids was first identified by virtue of antibacterial activity from the environmental DNA clones (Brady and Clardy, 2000). However, recently a new prevailing hypothesis for more plausible alternative of N-acyl amino acids function in chemical signalling was suggested based on the chemical similarities between N-acyltyrosines and N-acylhomoserinelactones which play role in chemical signalling in Gram-negative bacteria (Craig *et al.,* 2011). It has also been suggested that some environmental DNA-derived N-acyltyrosine synthase genes represent an expansion of the lipid biosynthetic machinery responsible for the synthesis of the bacterial type ornithine-containing lipids, which are structural lipids similar to phosphatidic acid (Geiger *et al.,* 2010).

2.4.5 Trehalose synthase and its role for viability at low temperatures

Grouped together with α-amylase, trehalose synthase was also detected in the metagenome. Trehalose synthase (TreS) catalyses the reversible interconversion of maltose and trehalose and has been shown recently to function primarily in the mobilization of trehalose as a glycogen precursor (Zhang *et al.,* 2011). Trehalose is an ubiquitous molecule that occurs in lower and higher life forms but not in mammals. It was previously believed to function solely as a reserve energy and carbon source in a manner similar to that of glycogen and starch (Elbein *et al.,* 2003). In industry, TreS has value as a simple biocatalyst for trehalose

production (Yue *et al.,* 2009). TreS was mainly studied in thermophiles and mesophiles, but seldom in cold-adapted microorganisms. Up to now, a few TreS genes from different bacterial species have been reported to be cloned, expressed and characterized, including *Pseudomonas stutzeri* CJ38 (Lee *et al.,* 2005b), *Pimelobacter* sp. R48 (Tsusaki *et al.,* 1996), *Mycobacterium smegmatis* (Pan *et al.,* 2004), *Arthrobacter aurescens* (Xiuli *et al.,* 2009), *Corynebacterium glutamicum* ATCC13032 (Wolf *et al.,* 2003), *Enterobacterhormaechei* that was isolated from Tibetan Plateau (Yue *et al.,* 2009) and some thermophilic strains (Koh *et al.,* 2003, Wei *et al.,* 2004).

The Ace Lake surface remains covered in approximately 2 m of ice for about 10–11 months (February/March–December) and usually (but not always) melts out during summer (January)(Gibson, 1999, Rankin *et al.,* 1999). Therefore, the cellular membranes of resident microbes are subject not only to rigidity but also to physical damage from ice-crystal formation during the freezing process (Deming, 2009). The presence of antifreeze protein activity in this lake has been demonstrated by Gilbert *et al.* (2004). In addition to antifreeze protein, the detection of TreS in this environment also suggested the role of trehalose as a cryoprotectant. Small molecular weight sugars were commonly used as cryoprotectants in the deep-freezer (-80°C) storage of microbes (e.g: glycerol), presumably providing a buffer between cells and ice crystals. The mechanism by which trehalose mediates tolerance to freezing or desiccation is not clear, but presumably involves a stabilisation of certain cell proteins and/or lipid membranes.

Another role of trehalose in cell protection against cold temperature was its effect in reducing oxidative damage. The solubility of gasses increases rapidly at low temperature especially for a dioxygen, which is a very reactive molecule. The dissolved O_2 concentration measured at the Ace Lake surface (5-11m) was ~13mg/mL (Lauro *et al.,* 2011). Ng *et al*. (2010), indicated that microorganisms in the Ace Lake environment were prone to oxidative stress and identified several oxidative stress proteins that may fulfil roles in oxidative defense in the 12.7 m Ace Lake metaproteome. Exposure of cells to H_2O_2 caused oxidative damage to amino acids in cellular proteins, and trehalose accumulation was found to reduce such damage presumably by acting as a free radical scavenger (Benaroudj *et al.,* 2001).

2.5 Conclusion

The analysis of the sequence-based search for hydrolases indicated the presence of the subtilase, lipases and GH13 genes throughout the lake. However, the matches for complete hydrolases genes were almost limited to samples of 0.1 μm mixolimnion. This result might be driven by at least two factors. The first is the advantage of the assemblies of combination Sanger and 454 pyrosequencing reads for 0.1 μm samples. The latter could be that the mixolimnion samples constitute of novel sequences that were inaccessible by sequence-based search techniques that was depending on the readily available HMM in Pfam database. The analysis of domain architecture of the sequences obtained, revealed the novelty of some of the sequences and inferred their physiological role in the native environment. The stringent sequence screening process depending on minimum ORF length, complete ORF and conserved catalytic sites might overlook some relevant sequences from the entire datasets. However, the functional screening of the library as described in the next chapter was performed to cover some of these potential gaps. Sequences from the 'lenient' group (without complete catalytic sites), were included in the functional screening process of the metagenomic library as described in the next chapter.

CHAPTER 3

Functional screening of hydrolases from the Ace Lake metagenomic clone library

3.1 Introduction

The increasing demand and value for novel biocatalysts by industries, e.g., pulp and paper, biofuel production and cosmetic and pharmaceutical products, has stimulated the exploration of various extreme environments, such as Antarctica, that possibly hold enormous varieties of cold-adapted enzymes (Lohan and Johnston 2005). The demand is supported by the advancement in the field of metagenomics that offers unique perspectives on the unculturable psychrophilic microorganisms and their biosynthetic products (Voget *et al.,* 2005, Green and Keller, 2006). As introduced previously, the culture-independent technologies were developed from the improvement made in DNA extraction and cloning directly from the environmental samples (Schmidt *et al.,* 1991, Handelsman, 2004). One of the critical steps in the metagenomic analysis is to screen for clones that contain target genes amongst a large number of clones in the metagenomic library. In general, labour intensive analyses of individuals or pool of clones within the library is often required for the activity-based screening procedure. The screening process normally involves analysis of several hundred thousand clones to detect a few (0.001-0.01%) functionally active clones (Henne *et al.* 2000; Heath *et al.* 2009; Berlemont *et al.* 2011).

Among the different classes of enzymes, hydrolases are of particular importance in the industry due to their broad substrate specificities, high stereo- and regioselectivities, cofactors-free activities, and some even stable in organic solvents (Faber, 2011). Cold-adapted proteases, amylases and lipases are of great significance in the biotechnology sector with applications such as in detergent, food and drinks, fermentation, textile, and pulp and paper industries as reviewed by Cavicchioli and colleagues (Cavicchioli *et al.,* 2011). This chapter describes the screening of the Ace Lake metagenomic clone library for protease, lipase, and amylase activity using agar-based assays. The library was previously constructed for classical shotgun sequencing purposes based on Sanger sequencing method. The clones

that used in the screening process are based on two groups. The first group was the targeted clones which contained the specific hydrolase genes identified in the Chapter 2. These clones were retrieved from the clone library through barcode identification. The second group consisted of ~20000 randomly-picked clones (not specifically picked based on the sequencing barcode nor identified as harbouring the gene of interest in the sequence-based search). Any activity detected on agar-based assays was subsequently confirmed by liquid enzymatic assay. An overview of the screening process is illustrated Figure 3.1.

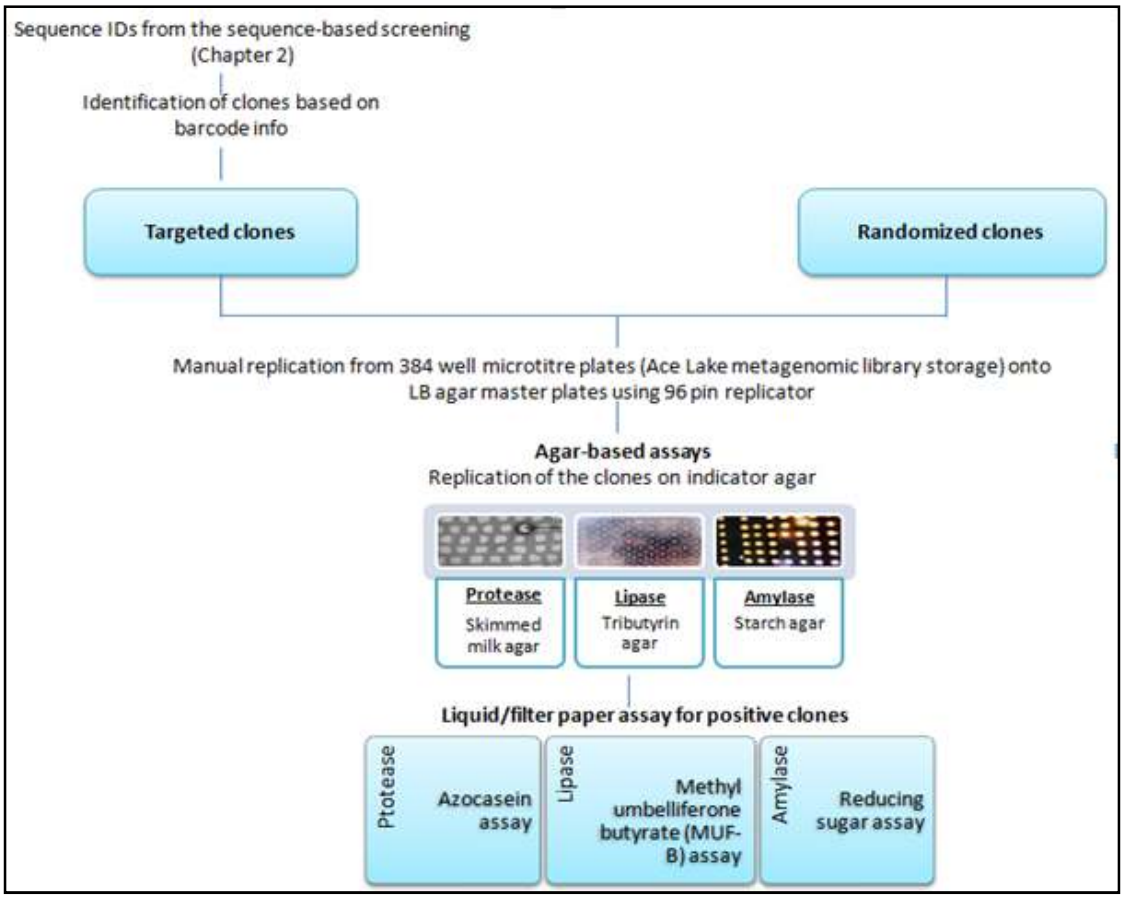


Figure 3.1: Flow chart of the enzyme screening process.

3.2 Materials and Methods

3.2.1 Library construction

Ace Lake metagenomic library construction was performed at the J. Craig Venter Institute (JCVI), in Rockville MD, USA (Rusch *et al.,* 2007), using a series of pHOS vectors and *E. coli*

Thunderbolt GC10 (Sigma, USA) as the host. The insert size ranged between 2-6 kb. All the clones were replicated and stored in 384 well microtitre plates with systematic barcode labelling that was used to track the clones with their respective sequence information in the Antarctic metagenome dataset.

3.2.2 Identification of the targeted gene from the sequence-based screening in the source clone library

The clone libraries in 384 well microtitre plates were stored in a hotel of five columns that each had a unique HOTEL_BARCODE and HOTEL COLUMN_BARCODE. Read ID that linked the clones to the assemblies that contained subtilase, lipase and amylase genes were used to retrieve the FASTA sequence for each clone. The individual sequence FASTA header contained the library ID (library name e.g ANTRC231), the well ID, which was called SEQUENCER_PLATE_WELL_COORDINATES (384 wells per plate, rows A to P, columns 1 too 24) and the GROWTH_BARCODE, which linked each sequence to the source clone plate. The GROWTH_BARCODE could be interchanged with the CONTAINER_BARCODE and was used to identify HOTEL_BARCODE and HOTEL_COLUMN_BARCODE from the spreadsheet (Appendix 3). Information from the FASTA header was linked to the information on the spreadsheet to identify each clone in the Ace Lake metagenomic clone library.

3.2.3 Agar-based functional screening of hydrolases from the Ace Lake metagenomic clone library

3.2.3.1 Randomised functional screening

~20000 randomly-pick clones (not specifically picked based on the sequencing barcode nor identified as harbouring the gene of interest in the sequence-based search) from the Ace Lake metagenomic clone library were replicated manually from the 384 well microtitre plates onto the Luria Bertani (LB) agar using a 96 pin replicator. After an overnight incubation at 30°C, the clones were replicated onto LB agar plates containing 2% (w/v) skimmed milk, 1% (w/v) potato starch (Lämmle *et al.,* 2007) and 1% tributyrin (v/v) (Torres *et al.,* 2003) to screen for protease, amylase and lipase activities respectively. All indicator plates, except for starch agar, were incubated at 30°C overnight to allow the cells to grow before being

transferred to 30, 25 and 4°C for 2 weeks. Amylase activity was detected by flooding the starch agar plates with Gram's iodine solution (3.3 $mgmL^{-1}$ I_2, 2.8 $mgmL^{-1}$ KI)(Lammle *et al.,* 2007) after incubation for a week at 30, 25 and 4 °C. Enzymatic activities were detected by the formation of a halo surrounding the positive clones on the indicator agar. In the screening process, recombinant strain harboring cold-adapted subtilisin-like gene of *Shewanella* sp. Strain AC10 which was designated as *E. coli* DH5α-pSapSh3 (Kulakova et al. 1999) was used as positive control for protease activity. *Pseudomonas aeruginosa* and *Bacillus amyloliquifaciens* from School of Biotechnology and Biomolecular Science (BABS) culture collection were used as positive control for lipase and amylase activities respectively.

3.2.3.2 Targeted functional screening

All targeted clones were selected from the library as described in section 3.2.2 and subjected to the screening process such as described in section 3.2.3.1.

3.2.4 Enzymatic assay

Enzymatic assays were performed using both cell-free supernatant and the cell extract of the cultures containing the clones of interest. The crude extracellular cold-adapted subtilisin expressed and secreted by of *E. coli* DH5α-pSapSh3 (Kulakova *et al.,* 1999) was used as positive control for protease activity. Crude cold-adapted amylase, Palkozyme CL L (Maps (India) Ltd) and porcine lipase (Sigma) were used as positive controls for amylase and lipase activities respectively.

3.2.4.1 Samples preparation

3.2.4.1.1 *Cell-free supernatant*

Cell cultures containing the positive clones were grown for 8, 16 and 24 hours according to the experiment requirement in 100 mL LB media. The cultures were centrifuged at 4000 g for 30 mins at 4°C to harvest the cell-free medium. The cell-free medium, referred to as cell-free supernatant, was used to identify the presence of extracellular enzyme. To concentrate the cell-free medium, ammonium sulfate precipitation was performed. Ammonium sulfate was dissolved completely in the 50 mL cell-free supernatant with gentle stirring to 75%

saturation. This step was done on ice to avoid protein degradation. The solution was then incubated overnight at 4°C. The amount of ammonium sulfate added was calculated based on the table in the Appendix D. After overnight incubation, the solution was centrifuged at 10000 g for 30 mins in a precooled rotor to precipitate the insoluble material. The pellet was resuspended in 4 mL of 50 mM Tris-HCl pH 7.5 buffer to yield a concentrated cell-free supernatant (25x), and kept at 4°C until further use.

3.2.4.1.2 *Crude cell extract*

The cell pellet from a 100 mL culture was resuspended in 1 mL of 50 mM Tris-HCl buffer (pH 7.5) and the cells were lysed by sonication. Sonication was performed on ice for 1 mins (0.5 s pulse on; 0.5 s pulse off; amplitude: 30%) using a digital Branson sonicator followed by centrifugation at 15000 rpm for 30 mins at 4°C. The supernatant was referred to as the crude cell extract.

3.2.4.2 Azocasein assay

The azocasein assay was used to detect protease activity in the cell-free supernatant and crude extract from the cultures containing the positive clones. The protease activity is determined by measuring the absorbance (405nm) of the azo-molecules released upon azocasein hydrolysis. 100 µL of protease sample was added to 0.5 mL of the substrate solution (50% w/v) urea, 0.6% azocasein powder, 0.2M Tris-HCl (pH 7-8)) and incubated at 25 °C. Reactions were terminated either after 30 mins or after overnight incubation, by adding 0.5 mL of 10% (w/v) aqueous trichloroacetic acid (TCA). The samples were centrifuged at 13000 rpm for 20 mins. The absorbance readings of the supernatant at 405nm were recorded against a reagent blank.

3.2.4.3 3,5-dinitrosalicylic acid assay

An assay based on the 3,5-dinitrosalicylic acid (DNS) assay for reducing sugar (Wood and Bhat, 1988), was used with modifications to measure amylase activity in the cell-free supernatant and crude extract of the positive clones. 100 µL of the amylase were added into 500 µL of 2% starch solution (pH 7.5) and incubated at 25°C for up to 15 hours. 500 µL of DNS reagent was added to the reactions. The assay tubes were incubated for five mins in

boiling water to stop the reaction and then allowed to cool to room temperature. The absorbance values at 540nm (after subtraction of the reagent blank) were recorded. The absorbance values were translated into maltose concentrations using a standard curve of absorbance of maltose at known concentrations. Reaction mixture without the sample was used as reagent blank.

3.2.4.4 MUF-butyrate assay

The MUF-butyrate filter paper assay was used to detect lipase activity in the crude extract of the cells containing the positive clones. 5 µL of crude extract was transferred onto a filter paper. 5 µL of 25 mM methyl umbelliferone butyrate (MUF-B) stock solution was spotted onto the crude extract, followed by ultra violet (UV) illumination of the paper (Prim *et al*. 2003). Samples with the lipolytic activities illuminated when exposed to UV light.

3.2.5 Zymography

Zymography was a second method used to detect lipase activity in the cell-free supernatant and crude extracts of the cultures containing positive clones. Samples were prepared using native sample buffer (5x), 10 uL concentrated cell-free supernatant or and 1 µl crude cell extracts (prepared as described in section 3.2.4.1.1 and 3.2.4.1.2 respectively). The whole sample was loaded into native polyacrylamide gel (4% stacking and 10% resolving gel). The gel was electrophoresed at 4°C using 1x electrophoresis running buffer system at 50 V until the dye reached the resolving gel. The voltage was increased to 100 V and stopped when the dye front reached the bottom of the gel.

The substrate for the lipase assay was freshly prepared by dissolving 2.5 mg of α-naphthyl acetate in 1 mL acetone. The volume was adjusted to 10 mL with 50 mM Tris-HCL (pH 7.5). A pinch of fast blue salt dye (Sigma) was added and the volume was adjusted to 50 mL by using similar buffer. The gel was then immersed in the α-naphthyl acetate solution in the dark for 10 mins. Porcine lipase (Sigma) was used as positive controls.

3.2.6 Sequence analysis

The Hawkeye analysis tools (Schatz *et al.,* 2007) were used to identify the read IDs from the metagenomic assemblies. In-house generated genbank files corresponding to genes of interest were retrieved to access the annotation of the predicted protein in the respective scaffold.

3.3 Results

3.3.1 Agar-based assay

Three types of indicator agars for detecting protease, lipase and amylase activities were used in the agar-based functional screening process. The screening resulted in low hit rates, 1:10000 for both protease and amylase (Table 3.1). The following sections describe the results obtained from the functional screening for protease, lipase and amylase activities.

Table 3.1**:** The results for agar-based screening of a metagenomic DNA library for different enzymatic activities.

Enzymatic activity	Indicator agar	No. of colonies tested	No. of positive clones	Hits rate
Protease	Skimmed milk agar	20140	2	1:10000
Amylase	Starch agar	20080	2	1:10000
Lipase/Esterase	Trybutyrin agar	20000	-	-

3.3.1.1 Agar-based screening for protease activities

Randomised functional screening of ~20000 clones for protease activity did not yield any positive clones. However, the screening of the 140 clones identified from the sequence- based search showed that two clones, N21 and D24, formed a halo on skimmed milk agar after two weeks incubation at 25°C (Figure 3.2). Only one clone, D24, showed consistent halo formation on skimmed milk agar when regrown on the media.

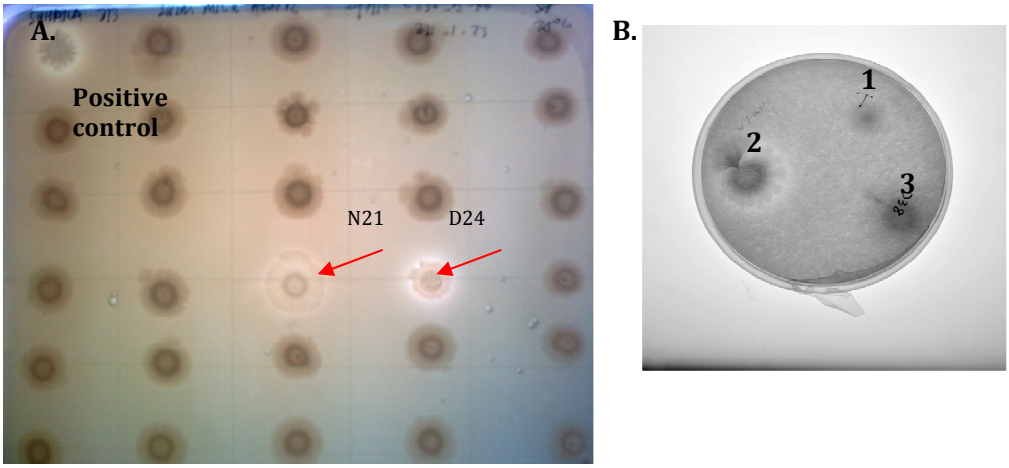


Figure 3.2**:** Results for agar-based screening for protease activity**.** A Proteolytic activity of clone N21 and D24 on skimmed milk agar screen. Positive control: *E. coli* DH5α pSAPSH3; B. Confirmation of the proteolytic activity of D24, 1 and 3: negative control, 2: D24.

3.3.1.2 Agar-based screening for amylase activities

The screening of the 64 targeted clones from the sequence-based search did not yield any positive clones. However, the randomised functional screening of ~20000 clones resulted with 23 clones with amylolytic activity. Of the 23 clones screened, two clones that showed the biggest halo formation on the starch-iodine media, i.e., D8 and D38, were chosen for further test (Figure 3.3).

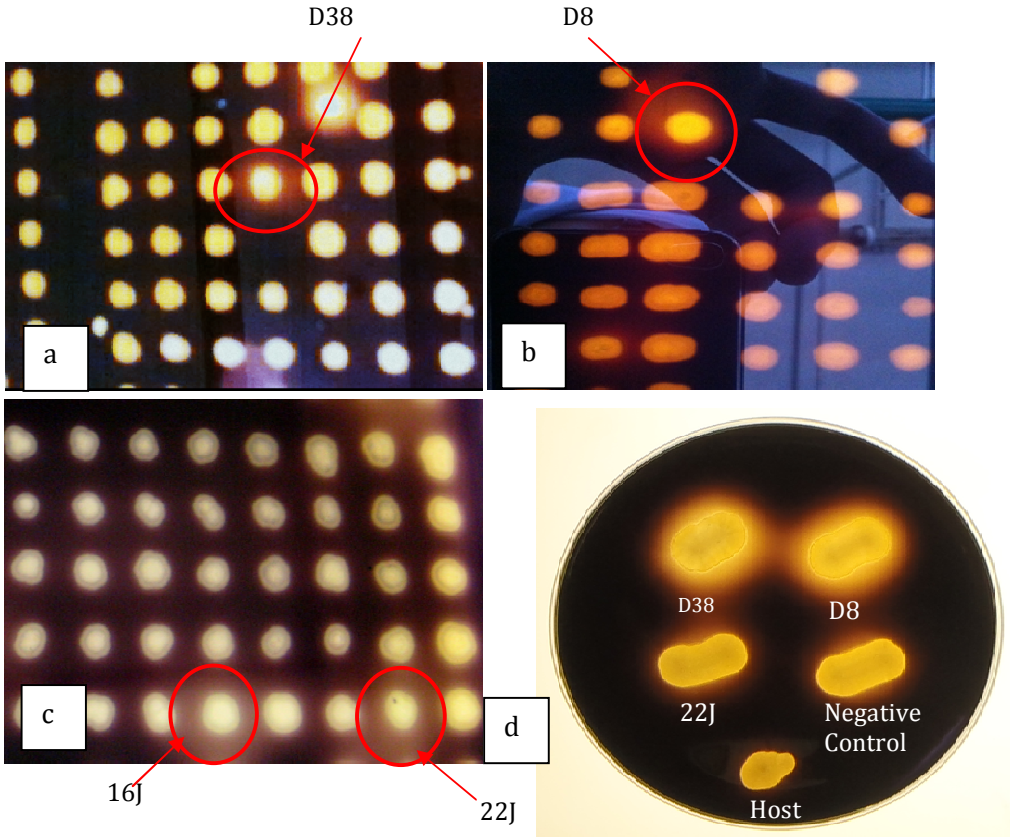


Figure 3.3**:** Results for agar-based screening for amylase activity on the starch agar. a, b & c: Positive clones from the randomised screening on starch-iodine agar. d: Confirmation of the amylase activities of clones D38 and D8 on starch-iodine agar.

3.3.1.3 Agar-based screening for lipase activities

After two weeks incubation at 25°C and 30°C, the screening for lipase activities on the tributyrin agar resulted with uniform halo formation surrounding all colonies (Figure 3.4). To detect whether this was due to lipase/esterase activity from the host cell, random clones were picked for further testing for lipase/esterase activities.

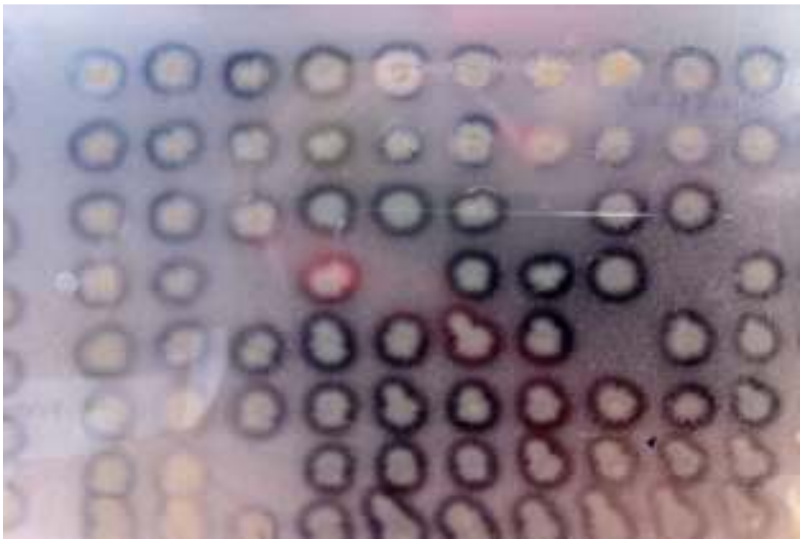


Figure 3.4: Trybutyrin agar assay after 2 weeks incubation at 25°C.

3.3.2 Liquid assay

3.3.2.1 Azocasein assay

The concentrated supernatant and the crude extracts from the cell containing positive clone, D24, that showed activity on the skimmed milk agar, were tested for protease activity using azocasein assay. However, no activity was detected. The absorbance reading at 405nm of the culture containing D24 clone and *E. coli* host strain was equivalent when both the extracellular and intracellular fractions were tested after 8, 16 and 24 hours of growth (Figure 3.5 (b & c)). There was no colour change of the substrate from red to yellow indicating the inability of the clone D24 to hydrolyse azocasein.Therefore, the growth profile of the cell containing D24 clone was established to see the possibility that the halo formation on the skimmed milk agar was due to cell lysis. The culture was grown in LB media at 25°C and compared to the growth profile of the *E. coli* host strain (negative control). Lysis of the cells containing D24 clone was indicated by the decline in cell density at early stationary phase as detected by taking optical density (OD) readings at 600nm (Fig 3.5(a)).

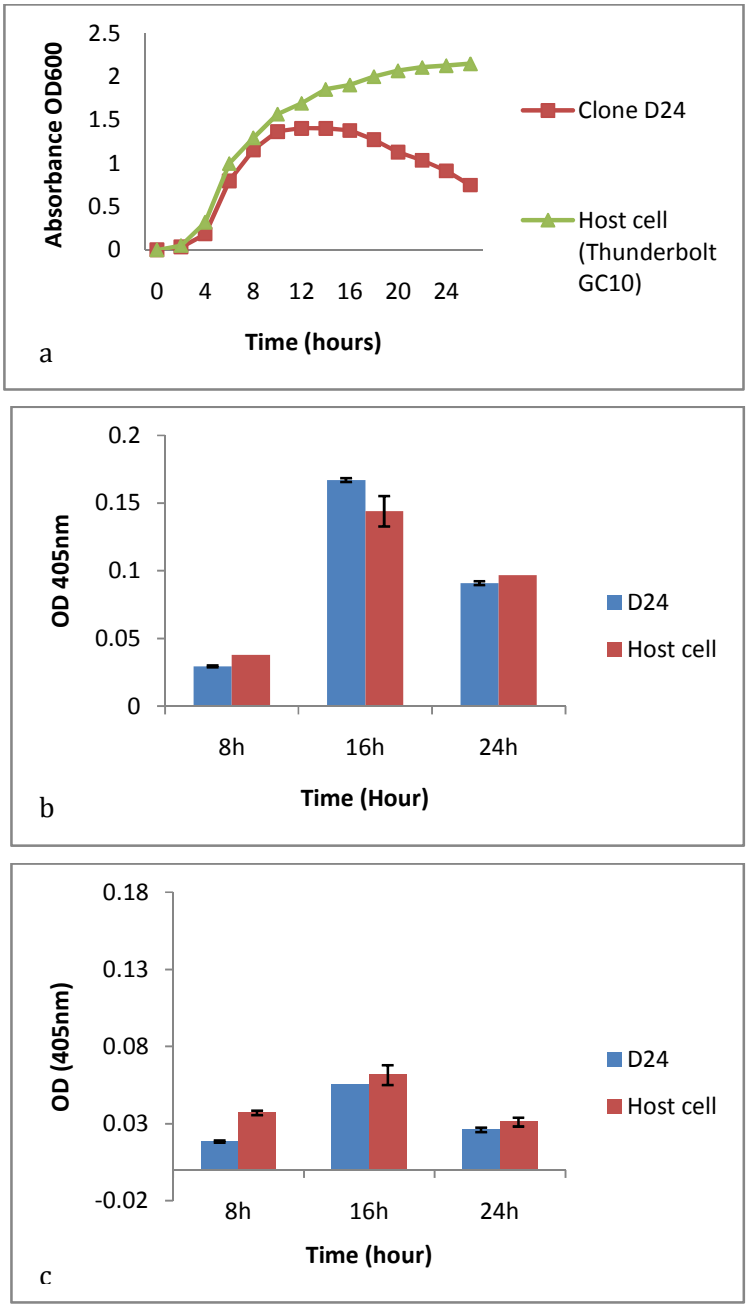


Figure 3.5: Confirmation data of the protease activity of clone D24 in azocasein assay. a: Comparison of the growth profiles of clone D24 and *E. coli* host cultures: b: Extracellular protease activity of the cell-free supernatant. c: Protease activity of the cell extracts.

3.3.2.2 3,5-dinitrosalicylic acid assay

Cells containing the two positive clones with amylolytic activity (D38 and D8) were grown in the LB media at 25°C to establish their growth profile and were compared to the negative control. The cell densities were comparable for both positive clones (D38 and D8) and negative control. OD readings at 600nm indicated no evidence of cell lysis for up to 28 hours (Figure 3.6). Intracellular and extracellular amylolytic activity of the clones was tested using DNS reducing sugars assay and compared to the negative control. The absorbance measured using a spectrophotometer at 540 nm is directly proportional to the amount of reducing maltose. Extracellular amylolytic activity of the clones was equivalent to the negative control suggesting the absence of any secreted amylase. Intracellular amylolytic activity of the clone D8 cell extract, performed at 30°C, was at least double that of the negative control at all time points of the enzyme assay (Figure 3.7). This result showed that clone D8 produced intracellular amylase.

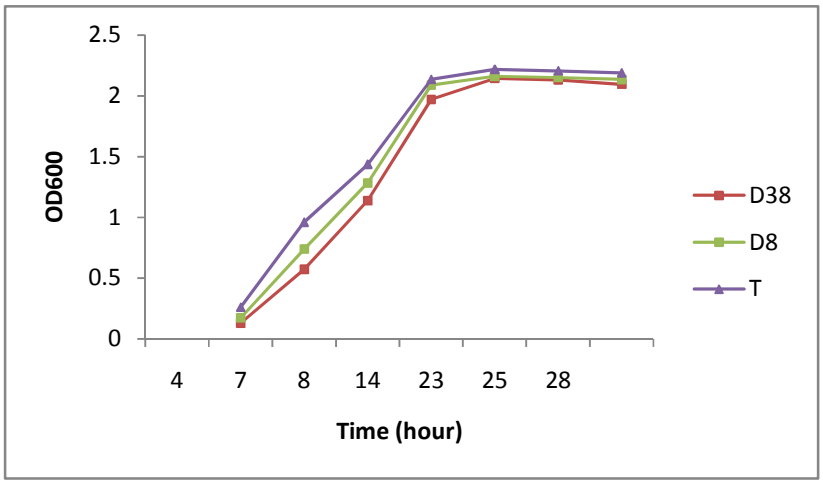


Figure 3.6: Growth curves (OD 650 nm) for the selected positive clones (D38 &D8) for amylase activity and the negative control (T).

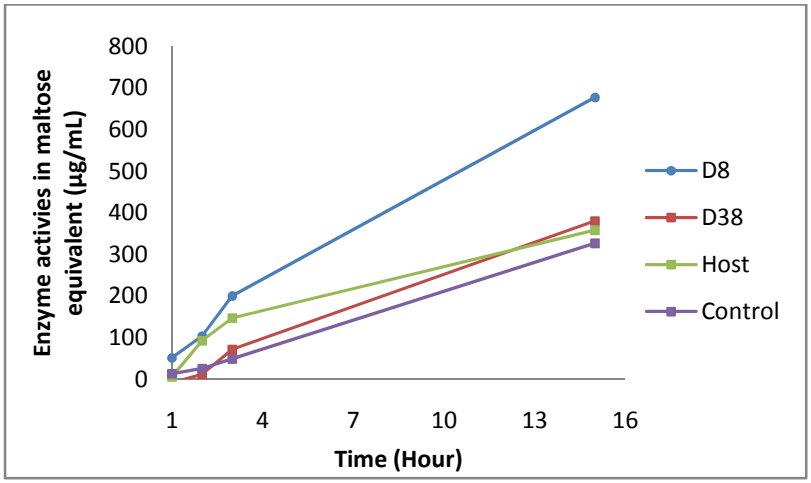


Figure 3.7**:** Enzyme activity assay (α-amylase) for intracellular extracts of the selected clones compared to the host and negative control using DNS assay (Wood and Bhat, 1988). One unit enzyme from the reducing sugar value release 1µg of reducing sugar (as maltose) from 2% starch solution in the respective time at 30°C, 20mM NaCl and pH 7.5 (unit/mL).

3.3.2.3 MUF-butyrate assay

The cell extracts of the selected clones from the initial screen were tested using the fluorescent indicator methyl umbelliferone butyrate (MUF-B) which detects lipase/esterase activities. Activities were detected in the host (Figure 3.8(ix)) and other selected clones. As expected no activity was detected for the negative control MUF-B (Figure 3.8(x)) which contained no cell extract.

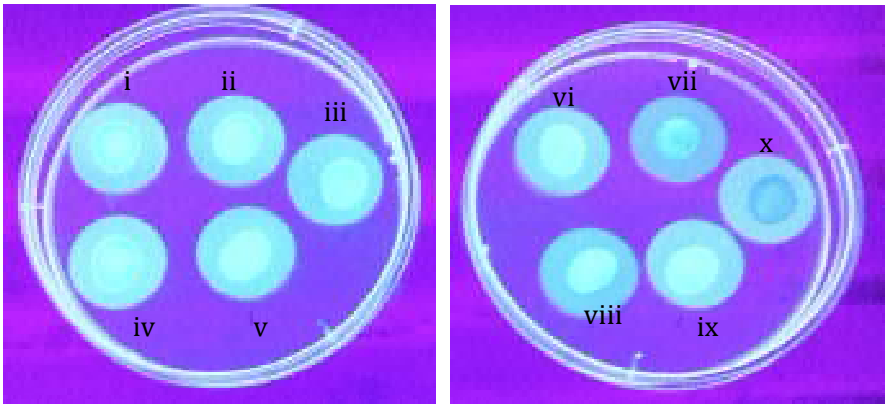
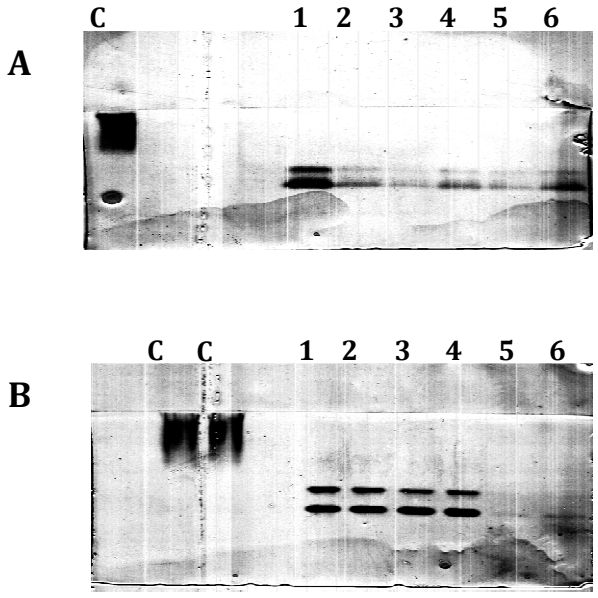


Figure 3.8: Result of the filter paper 4-MUF-butyrate assay. i: D20_M11, ii: D20_I7, iii: D20_F2, iv: 4ADNU_E22, v: D14_738_A7, vi: D36_1240_F9, vii: D14_738_A7, viii: D20_D24, ix: *E. coli* host strain, x: Negative control (4-MUF-butyrate only).

3.3.3 Zymography

To discriminate between clone and host generated esterase activity in the 4-MUF-butyrate assay, samples were subjected to Native PAGE zymography using α-naphtyl acetate as the substrate (Figure 3.9). Supernatant and whole cell extract from selected positive clones from the initial screen (L22, D22, F9 and K22) were compared to the host cell. The intensity of the lipase activity bands of the host was at least as strong as for any of the clones. The equivalent volume for each concentrated cell-free supernatant (10 µL) and the cell extract samples (1 µL), prepared as in the section 3.2.4.1.1 and 3.2.4.1.2 respectively, were loaded in the gel. The reduced band intensity for the clones indicates less protein was loaded than for the host strain. Overall, the zymogram did not reveal any new bands, or an intensity of bands that might indicate the clones were expressing recombinant active lipase.



A. C. Pig lipase control
1. *E. coli* host supernatant
2. *E. coli* host cell extract
3. Clone L22 supernatant
4. Clone L22 cell extract
5. Clone D22 supernatant
6. Clone D22 cell extract

B. C. Pig lipase control
1. *E. coli* host supernatant
2. *E. coli* host cell extract
3. Clone F9 supernatant
4. Clone F9 cell extract
5. Clone K22 supernatant
6. Clone K22 cell extract

Figure 3.9: Results of Native PAGE zymogram for lipase activity of selected clones, host and lipase positive control. α-napthyl acetate was used as substrate.

3.3.4 Sequence analysis

3.3.4.1 Analysis of scaffold-derived information for the clone with protease activity

The read ID for clone D24 that showed a halo on skimmed milk agar was 1113235614993. The read was included in a short scaffold of ~2kb, scf7180000121853 (Figure 3.10). The scaffold comprised of information from four pairs of reads from Sanger sequencing which were 1113289327225/1113297939409, 1113235614993/1113171833081, 1112328246771/112308938600, and 1112308881563/1112308926491). In-house annotation of this scaffold included two hypothetical proteins (163261385: 3...773) and (163261389: 1519...2190), and a 4-hydroxy-2-ketovalerate aldolase (163261387: 773....1507). The read 1113235614993 covered locus 1519...2190 thus indicating that the clone with protease activity was annotated as a hypothetical protein (Figure 3.11).

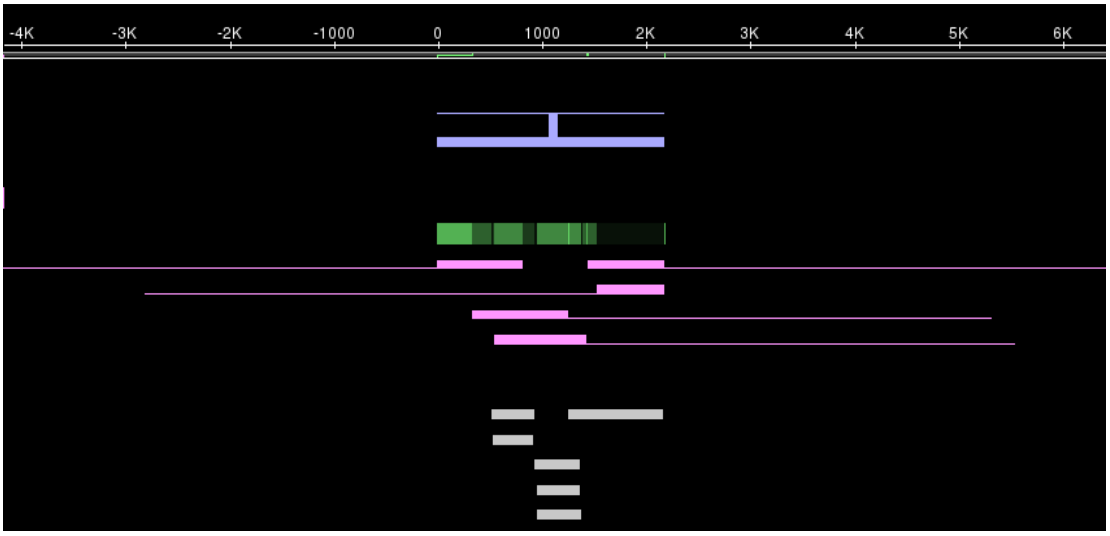


Figure 3.10 Overview of the read assembled in the scaffold 7180000121853.

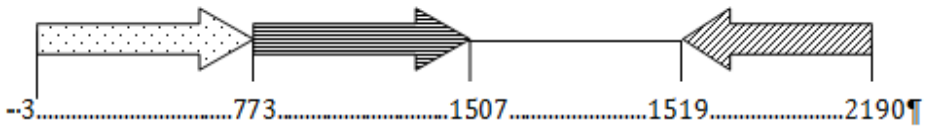


Figure 3.11: Schematic overview of the annotation in the scaffold 7180000121853. 3-773: hypothetical protein, 773-1507: 4-hydroxy-2-ketovalerate aldolase, 1507-1519: hypothetical protein.

Information from reads 1113235614993, 1112308926491, and 1112308938600 were found in a different long scaffold with a deep coverage, scf7180000134253, at locus 27733-33640 (Figure 3.12). From the in-house annotation, this locus range included ABC-type spermidine/putrescine transport system, ATPase component (163154514:27271…28296 forward), uncharacterized membrane-associated protein-like (28415…29056), hypothetical protein BcaDRAFT_1898 (163154515:29106…30194), UTP-glucose-1-phosphate uridyltransferase (163154516:30341-31249 reverse), peptidase S8 and S53 subtilisin kexin sedolisin (163154518: 31306…32616), 4-hydroxy-2-ketovalerate aldolase (163154520:32629…33363)(Figure 3.13). Specifically, read 1113235614993 sequence information covered locus 27733…32733 thus confirming the annotation as subtilisin instead of hypothethical protein in the scaffold 7180000121853.

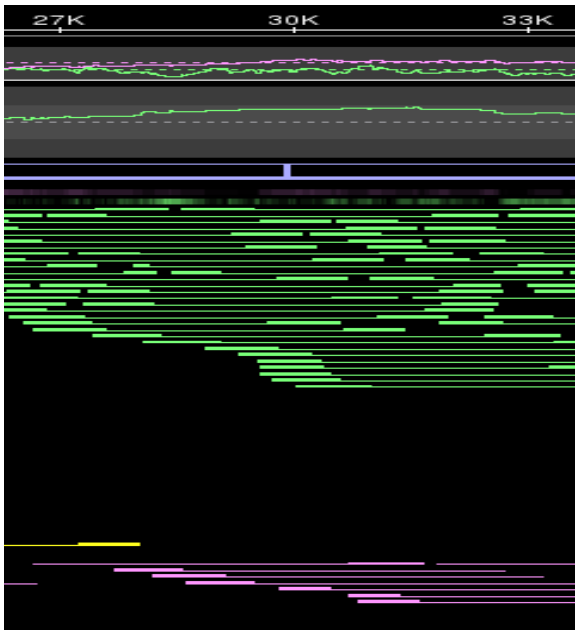


Figure 3.12 **:** Graphical overview of the reads assembly of part of the scaffold 7180000134253.

Figure 3.13**:** Schematic overview of the annotation in the scaffold 7180000134253 at locus 27271 towards 33363. 27271-28296: ABC-type spermidine/putrescine transport system, ATPase component, 28415-29056: uncharacterized membrane-associated protein-like, 29106-30194: hypothetical protein BcaDRAFT_1898, 30341-31249: UTP-glucose-1-phosphate uridyltransferase, 31306-32616: peptidase S8 and S53 subtilisin kexin sedolisin, 32629-33363: 4-hydroxy-2-ketovalerate aldolase.

3.3.4.2 Analysis of the clone with the amylolytic activity

The clone D8 that showed positive amylase activity in the agar-based screening and DNS assay was derived from randomised functional screening. The sequencing information of this clone was retrieved using the container plate ID and the sequencer well coordinate. The clone was screened from well D8, plate DD0004A4BNU with barcode BB0227127AB from ANTRC229-G-01-3-4KB Antarctica 2007 Environmental 0.1 um marine filtrate sequencing library. Using the barcode, the pair of reads 1112075940101/1112079329003 were retrieved. Further analysis found that both reads were not included in any assembly and were not annotated. The clone needs to be sequenced to identify the gene that encoded this starch degrading function.

3.4 Discussion

3.4.1 Functional screening of the metagenomic library

Results from the sequence-based search of the Ace Lake metagenomic dataset previously described in the Chapter 2, have confirmed the presence of hydrolase gene sequences. The readily available clone library of the Ace Lake metagenome allowed functional screening of the library for potentially cold-adapted hydrolases using agar-based assay. Information from the sequencing data was used to identify the clones that contained the hydrolase genes (targeted clones). Arguably, the clones were selected on the basis of the similarity of their catalytic domain to the available sequence in the public databases thus limiting the success of finding completely novel enzyme. However, the clones still have the potential to exhibit novel properties since the genes have been isolated from a pristine Antarctic cold-adapted environment that is already known to be the source of biotechnologically useful enzymes and compounds (Bowman *et al.,* 2005). In fact, a sequence-based screening approach was able to retrieve novel enzyme sequence, as demonstrated for polyketide synthase (Seow *et al.,* 1997) and xylose isomerase (Parachin and Gorwa-Grauslund, 2011).

The functional screening of the metagenomic library was further enhanced by randomised screening. This approach has proven to successfully reveal novel enzymes that independently of their similarity to previously known genes (Steele *et al.,* 2009). Clones were picked randomly from the library and subjected to the agar-based screening process together

with the 'targeted clones'. Protease, lipase and amylase activities were assayed on agar plates supplemented with subtrates, i.e., skimmed milk, tributyrin and starch, respectively. The positive clones were identified through visual screening for the appearance of clear zone (halo) on the agar plate. This screening method was utilized based on its advantageous such as low cost (relatively), did not require any special devices and can be performed at high-throughput (Uchiyama and Miyazaki, 2009). As discussed previously in Chapter 1, this method has proven successful in identifying novel enzymes from the metagenomics libraries. However, one of the limitations of this method was the faint positive signals (Uchiyama and Miyazaki, 2009), which could be the reason for the low hit rates in this study.

3.4.2 Low hit rates of hydrolase activity in the Ace Lake metagenomic library.

The agar-based screening process was performed at temperatures below 37°C, i.e., 30, 25 and 4°C. Previous studies have shown that, typically, the specific activity of the cold-adapted enzymes is higher than that of their mesophilic counterparts at temperatures of approximately 0-30°C (Gerday *et al.,* 2000). In fact, some cold-adapted recombinant enzymes might not fold properly in the mesophilic host, such as *E. coli*, or might be partially inactivated as a result of the high temperature (>30°C) used for their expression (Gerday *et al.,* 2000). For example cold-adapted triose phosphate isomerase expressed in *E. coli* at 37°C, has a specific activity that is four times lower than when the *E. coli* are grown at 27°C (Alvarez *et al.,* 1998).

Despite of the effort to screen for hydrolase activity at temperature below than 30°C, the agar-based functional screening in the Ace Lake metagenomic library resulted with low hit rates of 1:10000 for both protease and amylase activity (Table 3.1). A clone positive for lipase activity was not detected. The low detection rate of positive clones has been previously reported in the metagenomic screening studies of samples from the cold environments. For example, Heath *et al.* (2009) screened an Antarctic desert soil metagenomic library and found three clones with esterase activity out of 100,000 clones, Sharma and colleagues detected one amylase clone in the library of 350,000 Northwestern Himalayas soil clones (Eisen *et al.,* 2006) and only one xylanase clone was detected from 5,000,000 clones from a dairy farm wastewater metagenomic library (Lee *et al.,* 2006).

According to Uchiyama and Miyazaki (2009), the probability (hit rate) of certain genes depends on multiple factors that are inextricably linked to each other such as the host–vector system, the size of the target gene, its abundance in the source metagenome, the assay method, and the efficiency of heterologous gene expression in a surrogate host. The low detection of hydrolase active clones from both randomised and targeted screening approach is indicative of the difficulties of successfully expressing environmental DNA in a heterologous host. Some of the contributing factors to this problem are the inability of the *E. coli* transcriptional system to recognize promoter regions of environmental DNA (Gabor *et al.,* 2004), codon bias (Kane, 1995) and incorrect protein folding and transport (Baneyx and Mujacic, 2004).

While the hit rate of clones with hydrolase activity was low, the positive clones turned out to exhibit low enzymatic activity. This condition might be explained by the type of vectors being used in the construction of this library. Undeniably, the main purpose of the Ace Lake metagenomic library construction for shotgun sequencing reflected the type of vector and host chosen for the library. The library was constructed using a low copy number (~20 copies per cell) plasmid (pHOS-derivative of pBR194c) and the strong promoters oriented towards the cloning site of the vector were eliminated to ensure the construction of high quality random plasmid library (Rusch *et al.,* 2007).

With this inherent factor of the clone library, it was reasoned that if the screening proved successful, those genes could eventually be expressed in another expression system. In fact, irrespective of the low copy number and elimination of the strong promoter, the clones should be able to express any of the genes that were cloned together with their own promoter if recognised by the *E. coli* gene expression system.

3.4.3 The pitfalls of using skimmed milk and tributyrin agar-based assay to detect proteolytic and lipolytic activity in the metagenomic clone library.

In this project, one of the clones showed a halo formation on skimmed milk agar during the functional screening. However, this clone did not show activity in the azocasein assay. From the growth profile of the clone, it was likely that the halo formation was due to the release of

cell content following cell lysis. None of the annotations in the bioinformatic analysis of the clone were related to toxicity. Apart of annotation as subtilase, other annotation includes 4-hydroxy-2-ketovalerate aldolase, ABC-type spermidine/putrescine transport system, ATPase component, uncharacterized membrane-associated protein-like, glucose-1-phosphate uridyltransferase and hypothethical proteins.

The use of skimmed milk agar to screen for protease activity is common (Jones *et al.,* 2007). However, there was a report by Jones and colleagues (2007) that skimmed milk agar was not suitable for the screening of gut microbiome metagenomic library for protease activities. They found that the halo formation on skimmed milk agar was due to acid production from glycoside hydrolase activity instead of protease activity. However, using a similar screening approach, Waschkowitz and colleagues (2009) were able to isolate two positive clones with strong proteolytic activity from the metagenomic library of mixed sediment and soil samples. These findings, suggested that skimmed milk agar is still a versatile selection medium for protease producing clones in metagenomic studies. Nevertheless, in order to resolve the issues of possible false positives using skimmed milk agar, Morris and colleagues studied the used of Valio™ lactose-free milk agar, and found that it is an effective and robust agar for correctly identifying proteases by way of distinct zones of clearing around a bacterial colony (Morris *et al.,* 2012). The study had confirmed the inability of glycoside hydrolase positive metagenomic clone to clear the lactose-free milk agar. Morris and collegues (2012) suggested the use of lactose and fat free milk agar instead of skimmed milk agar as a more reliable and efficient for future screening of metagenomic libraries for protease activity

Even though tributyrin is commonly used to screen for lipase/esterase producing clones, this project faced multiple occurrences of false positive results when tributyrin media was used. All clones grown on tributyrin agar showed halo formation after 2 weeks incubation at 25°C. A similar issue was previously reported by (Litthauer *et al.,* 2010). However in contrast to skimmed milk agar, the reason for the presence of false positives on tributyrin agar was unclear (Litthauer *et al.,* 2010). In the context of false positive results during the screening of Ace Lake metagenomic library, it was expected that the halo formation on tributyrin agar was due to the esterase activity of the host. This argument was supported by the detection of esterase activity in the host using the 4-MUF-B filter paper

assay and zymography using α-naphthyl acetate as the substrate. Previously, there were genomic sequencing efforts that revealed *E. coli* has at least 13 genes encoding acetyltransferase or esterase-like activity (Blattner *et al.,* 1997). *E. coli* also had significant cytoplasmic esterase activity towards 1,1-O-isopropylideneglycerol (IPG) esters encoded by the YbfF gene (Godinho *et al.,* 2011).

Since tributyrin had been successfully used to screen for lipase in various metagenomic studies (Elend *et al.,* 2007, Kim *et al.,* 2009, Wei *et al.,* 2009), the duration of experiments and the level of expression of recombinant lipase must be considered as possible explanations for the formation of halo on the tributyrin agar. It appears that a tributyrin agar-based assay is a feasible method for detection of lipolytic activity in recombinant clones that are able to express a high level of lipase/esterase after incubation of less than 2 weeks. Notably, there are no reports about this specific limitation regarding incubation time for using tributyrin in the functional screening of the metagenomic libraries.

3.5 Conclusion

In conclusion, the targeted functional screening of the Ace Lake metagenomic library using information from sequence-based screening did not isolate any clones with hydrolase activity. However, the randomised functional screening was able to detect at least one clone with starch degrading activity. The clone that showed relatively high amylase activity compared to other clones in the reducing sugar test would be a potential candidate for further cloning and expression in a better expression system. However, since the clones lack any sequence annotation for α-amylase or glycosyl hydrolase, the gene responsible for the starch degrading activity needs to be identified, for example by primer walking prior to the cloning process.

Despite the wealth of different biotechnologically important enzyme activities derived from numbers of metagenomic studies, this project began with very little available information on protease. The availability of Ace Lake metagenomic sequences datasets and the DNA clone library resources allow the selection, amplification and over expression of subtilase genes (protease) in *E. coli*. These approaches are discussed specifically in Chapter 4. In Chapter 5, a comprehensive bioinformatic analysis was performed to explore the relative

abundance of subtilase and other peptidase genes in the Ace Lake metagenome and compared to another two potential sources in the neighbouring site, i.e., Organic Lake and Southern Ocean.

CHAPTER 4

Cloning and expression studies of the selected subtilase genes from the Ace Lake metagenomic sequencing library

4.1 Introduction

E. coli expression systems remain one of the most attractive choices for overexpression and protein purification because of the ability of the organism to grow rapidly and to high density using inexpensive substrates, its well-characterized genetics and the availability of an increasingly large number of compatible cloning vectors and mutant strains (Baneyx, 1999, Francis and Page, 2001, Yin *et al.,* 2007). A considerable amount of effort has been directed at improving the performance and versatility of this microorganism in order to obtain high expression levels of active recombinant protein (Baneyx 1999; Baneyx and Mujacic 2004). Despite its many advantages and widespread use, there are also drawbacks for using *E. coli* as an expression system. The probability of successfully expressing a soluble protein decreases considerably as the molecular weight of the protein increases above 60 kDa. Some proteins do not express in the soluble form which may be due to the fact that the protein is not modified or folded properly, or some may precipitate to form inclusion bodies (IB)(Gräslund *et al.,* 2008). The likelihood of misfolding is increased by the routine use of strong promoters and high inducer concentrations that can lead to product yields exceeding 50% of total cellular protein. A second factor contributing to IB formation is the inability of bacteria to support all post-translational modifications that a protein may require to fold correctly (Baneyx and Mujacic 2004). The formation of IB can be a significant hindrance in obtaining soluble, active protein in recombinant protein production. However, in some cases, IB are advantageous because they are resistant to proteolysis, easy to concentrate by centrifugation, minimally contaminated with other proteins, and, with some effort, the protein can be refolded to an active, soluble form (Gonzalez-Montalban *et al.* 2007; Brondyk 2009).

In order to increase the efficiency of the refolding process, IB need to be washed free of contaminants (such as cell debris, and degraded DNA) with detergents such as Triton X-

100, Sarkosyl or lower molar concentrations of denaturants, such as 1 M urea. Generally, to achieve high protein solubility, the IB are solubilized using high concentration of denaturants usually 8 M urea and 4-6 M GdnHCl or low concentration of detergent such as SDS and 0.3-2 % Sarkosyl, before being refold in the specific refolding buffer (Rudolph and Lilie 1996; Tsumoto *et al.* 2003; Burgess 2009). The components of the refolding buffer vary widely in pH, ionic strength, redox condition and ligand, depending on the protein of interest with (Cabrita and Bottomley, 2004). The refolding is achieved via dialysis, dilution or on-column refolding (Burgess, 2009b). The process of refolding with effective removal of any denaturant is not a trivial process as it involves a competing reaction of misfolding and aggregation (Rudolph and Lilie, 1996, Cabrita *et al.,* 2006).

Apart from obtaining active forms of protein from IB via refolding, there is another method known as fusion protein technology which focus on elimination of IB during the recombinant protein expression process (Sorensen and Mortensen, 2005). The fusion protein technology use tags, such as *Shistosoma japonicum* glutathione S-transferase (GST), *E. coli* maltose-binding protein (MBP) and *E. coli* N utilization substance A (NusA), for overcoming IB formation and simultaneously for increasing soluble recombinant protein expression levels (LaVallie *et al.,* 1993, Davis *et al.,* 1999, Kapust and Waugh, 1999). The NusA-fusion protein system (Davis *et al.,* 1999) that is commercially available from Novagen, was used in this project. This system fuses the recombinant gene with a high molecular weight NusA protein (495 aa) for higher level expression of the recombinant protein. The system has proven to be successful to increase the solubility of recombinant proteins (De Marco *et al.,* 2004). This high molecular weight fusion protein was also used in several high-throughput expression studies (Shih *et al.* 2002; Cabrita *et al.* 2006) and resulted in increased expression of soluble recombinant protein.

Recombinant subtilase has been successfully expressed as an extracellular enzyme using the *E. coli* expression system (Kulakova *et al.,* 1999, Arnórsdóttir *et al.,* 2002). The most studied subtilisin proteases were subtilisin E, from *Bacillus subtilis* (Stahl and Ferrari, 1984), subtilisin BPN', from *Bacillus amyloliquefaciens* (Vasantha *et al.,* 1984) and subtilisin Carlsberg, from *Bacillus licheniformis* (Jacobs *et al.,* 1985). Extracellular subtilisin proteases are synthesized in a precursor form called pre-pro-subtilisin, in which a presequence (signal peptide) and a prosequence (propeptide) are attached to the N-terminal region of the mature

domain (Siezen and Leunissen, 1997). They are secreted outside the cell as a pro-subtilisin with assistance of a signal peptide. Upon completion of folding, the propeptides are autoproteolytically removed because they are not necessary for the activity or stability of their cognate folded enzymes (Shinde and Inouye, 2000). Furthermore, propeptides function as a potent competitive inhibitors of the enzymatic activity (Li *et al.,* 1995). The propeptides were also termed intramolecular chaperones due to their covalent attachment to the proteins that they help to fold (Inouye, 1990). The mature domains alone are folded into an inactive form with molten globular structure in the absence of the propeptide (Shinde and Inouye, 1995).

In this chapter the cloning and overexpression of the selected subtilase genes in *E. coli* is described. The genes were identified from the sequence-based screening of the Ace Lake metagenomic datasets discussed in Chapter 2. One of the genes that were identified as extracellular subtilase from the primary sequence, was expressed in the insoluble fraction. The attempt to obtain active enzyme by using refolding methods and by fusioning it with high molecular weight NusA protein is described. The proteolytic activity of one of the subtilase proteins that expressed in the intracellular soluble fraction is also reported.

4.2 Materials and methods

4.2.1 Physico-chemical analysis

The amino acid sequences of subtilase was submitted to the online tool "ProtParam", accessed via the ExPasy proteomics server (http://au.expasy.org/tools/protparam.html) for calculation of the molecular weight, theoretical pI, the molar extinction coefficient and grand average of hydropathicity (GRAVY). A hydropathy plot was prepared according to Kyte and Doolitle at (http://gcat.davidson.edu/DGPB/kd/kyte-doolittle.htm). Signal peptide prediction was done using SignalIP (Bendtsen *et al*. 2004).

4.2.2 Structural modelling

Three dimensional structures of Subt9195 and Subt8715 were determined computationally using I-TASSER webserver (Roy *et al.,* 2010, Zhang, 2008)

(http://zhanglab.ccmb.med.umich.edu/I-TASSER/), which gave the best protein models in the Critical Assessment of Structure Prediction (CASP7, CASP8, CASP9, and CASP10). CASP, is a community-wide, worldwide experiment for protein structure prediction that is held biennially.

4.2.3 Cloning of subtilase genes

4.2.3.1 PCR amplification of subtilase genes from Antarctic clone libraries

Polymerase chain reaction (PCR) was performed in a 50 µL volume with a final concentration of 0.02 U Taq DNA Polymerase, 1x ThermoPol Taq Reaction buffer, 200 µM deoxynucleotide solution mix, 0.1 µM each of forward and reverse primers (Table 4.1), 1.5 mM $MgCl_2$ and 0.1-1 ng/mL plasmid DNA. The PCR cycling parameters were as follows: initial denaturation phase at 95°C for 60 s, and followed by 30 cycles of denaturation, annealing, and extension at 95°C for 30 s, 54°C for 60 and 72°C for 2 mins, respectively. Amplification was finalised with an extension step at 72°C for 2 mins before being kept at 4°C until further use. Thermal cycling was performed in the MyCycler thermal cycler (Bio-Rad).

4.2.3.2 Preparation of vectors and insert for cloning process

4.2.3.2.1 Purification of PCR product from agarose gel

PCR products were verified by gel electrophoresis and purified using QIAquick Gel Extraction Kit (Qiagen), as per manufacturer's recommendations. In brief, the gels were sliced, dissolved, and the PCR products were subsequently captured and purified with silica-membrane based spin columns. Purified PCR products were eluted and stored at 4°C until further use.

4.2.3.2.2 Restriction digests of vectors and purified PCR product

Digestions of the purified PCR products and vectors for cloning were set up in volumes of 50 µL using restriction enzyme and buffer as recommended by manufacturer; New England Biolabs (NEB). Restriction digests were performed for 2 hours at 37°C. The digested PCR products were purified from the reaction mix using a PCR purification kit (Qiagen).

4.2.3.3 Ligation of vector and insert

The digested, purified PCR products were then ligated into the desired pET vector using T4 DNA ligase (NEB). The ligation was performed overnight at 16°C and stored at 4°C until further use.

Table 4.1: Primers used in PCR amplification of selected subtilase genes. Restriction sites are shown in italics and underlined.

Primer Id	Primer Sequences (5' ⟶ 3')	Restriction Site	Vector
Subt_9195-F	GACTA*CCATGG*ATATGCTGAAGAAAAACCCG	Nco 1	pET28b
Subt_9195-R	TAGA*GCTAGC*CTACCGCGTTCGCC	Nhe 1	
Subt_8715-F	TACTA*CCATGG*ATGTGACCGACCTGGTCG	Nco 1	
Subt 8715-R	ATCTG*GCTAGC*CTAGTTCGCGGTGACGG	Nhe 1	
Subt_5372-F	GACTA*CCATGG*ATGTGCAGTTAGCCGCCGGG	Nco 1	
Subt_5372-R	CGTAGA*GCTAGC*TTACTTTCGTATTTTTCTACCC TGGCCCT	Nhe 1	
Subt_4518-F	GACTA*CCATGG*ATTTGACTGACAGCTTTAAGGG GAT	Nco 1	
Subt_4518-R	CGTAGA*GCTAGC*CTATTTGGTAACTACCGTTTT AGAAAAAGC	Nhe 1	
Subt_9195_NusA-F	GAG*GAGCTC*GTATGCTGAAGAAAAACCCGATCC TTG	Sac 1	pET43a
Subt_9195_HisC-R	TAGA*CTCGAG*CCGCGTTCGCC	Xho1	

4.2.4 Competent cell preparation

A volume 0.5 mL of overnight culture was inoculated into 50 mL of LB broth. The inoculated broth was cultured at 37°C with shaking at 200 rpm until OD readings at 600nm reached 0.4-0.6. The culture was subsequently centrifuged at 4000 g for 10 mins at 4°C. The cell pellet was resuspended in 1 mL, and further with another 24 mL of 0.1 M of cold sterile $CaCl_2$. The cell suspension was incubated on ice for 30 mins before being centrifuged at 4000 g for 10 mins at 4°C. The cell pellet was then resuspended in 2.5 mL of cold sterile 0.1 M $CaCl_2$ 20% (v/v) sterile glycerol, aliquoted into 50 µL volumes, flash frozen in liquid nitrogen and stored at -80°C until further use. The competent cell strains of *E. coli* TOP10, BL21 (DE3) and Rosetta 2 (DE3) were successfully prepared using this method. Rosetta 2 (DE3) strain required the addition of the antibiotic chloramphenicol during culturing.

4.2.5 Cell transformation

1-2 µL of plasmid (10-100 ng) or 2-5 µl of ligation mixture was added to the 50 µL of competent cells that were already thawed on ice. The cells were mixed by tapping the tube gently and incubated on ice for 30 mins. The cells were heat shocked at 42°C for 90 sec and placed immediately on ice for five mins. 950 µl of LB broth was added to the cells before being incubated at 37°C for 1 hour with shaking at 150 rpm. Subsequently the cells were centrifuged at 13000 rpm for 1 min and the pelleted cells were resuspended in 100 µL of sterile LB broth. 50 µL of the cell suspension was plated on LB agar containing appropriate antibiotics.

4.2.6 Identification of positive clones

4.2.6.1 Colony PCR

Colonies were picked with a sterile pipette tip, which was touched on the interior of a PCR tube that contained the PCR mix. The same tip was patched onto the LB agar plate containing the appropriate antibiotic to maintain the clone. A PCR was performed as described in section 4.2.3.1. After amplification, the PCR product was electrophoresed on 1 % agarose gel and visualised using ethidium bromide.

4.2.6.2 Restriction digest test

For confirmation of positive clones identified by colony PCR, or to search for positive clones in cases where the basic screen failed to give reliable results, a restriction digest was performed on the plasmids isolated from the positive colonies. Plasmids were digested with restriction enzymes as described in section 4.2.3.1.2 used in the cloning procedure. The resulting fragments were electrophoresed on 1 % agarose and visualised using ethidium bromide.

4.2.7 DNA sequencing

4.2.7.1 Preparation of sequencing reaction sample

Samples for DNA sequencing were prepared in 20 µL total volume containing 1 µL of BigDye terminator V3.1, 3.5 µL of 5x buffer, 100-500 ng plasmid or 100-500 ng plasmid, 3.2 pmol sequencing primer and nuclease-free water. The sequences of the sequencing primer are illustrated in Table 4.2. The thermocycling parameters for sequencing reactions were as follows: 96°C for 10 sec, 50°C for 5 sec, and 60°C for 4 sec. These cycles were repeated for 25x before put on hold at 4°C until ready to be to be purified.

Table 4.2: Primers used for sequencing

Primer Id	**Primer Sequences**	**Recombinant Construct**
T7promoter-F	TAATACGACTCACTATAGGG	pET28b_Subt9195, pET28b_Subt8715, pET28b_Subt5372
T7terminator-R	GCTAGTTATTGCTCAGCGG	
S tag 18mer	GAACGCCAGCACATGGAC	pETt43a_Subt9195
Colidown R	TTCACTTCTGAGTTCGGCATGG	

4.2.7.2 Purification of sequencing reaction sample

Ethanol/EDTA precipitation method was used to remove unincorporated dye-labelled terminators. The entire volume of the sequencing reaction was transferred into 1.5 mL microtubes and added with 5 µL of 125 mM EDTA and 60 µL of 100% ethanol. The tubes were vortexed briefly and incubated at room temperature for at least 15 mins. Samples were then centrifuged at 14000 g for 20 mins and immediately the supernatant was aspirated and discarded completely. 160 µL freshly prepared 70% ethanol was added to the samples, vortexed briefly and centrifuged again at 14000 g for 10 mins. After the supernatant was aspirated and discarded the previous steps were repeated using 80 µl of freshly prepared 70% ethanol. Finally the samples were dried completely in a Speedvac and stored at -20°C in the light protected tube.

Samples were sent to the Ramaciotti Centre, University of New South Wales for sequencing. The chromatogram was assessed visually using sequence scanner software version 1.0 (Applied Biosystem) to identify any miscalled nucleotides.

4.2.8 Heterologous expression of recombinant proteins

The cloned genes were expressed in *E. coli* BL21 (DE3) and Rosetta 2 (DE3). The cells were grown in 50 mL LB medium containing appropriate antibiotics at 37°C with shaking at 200 rpm until OD 600nm reached–between 0.8-1.0. Cultivation temperature was then lowered to 30-16°C, and protein expression was induced with 100 µM to 1 mM of isopropyl-β-D-thiogalactopyranoside (IPTG). Cultures induced at 16 and 25°C were incubated overnight, while those induced at 30°C were incubated for 7 hours before harvesting**.**

4.2.9 Protein sample preparation

4.2.9.1 Preparation of extracellular protein fraction

The extracellular protein fraction was prepared to identify the secreted protein by the cell. 1 mL of bacterial cultures was harvested by centrifugation at 5000g for 20 mins at 4°C. The supernatant (medium) was collected. Proteins in the medium were precipitated by addition of 100 µL of 100% TCA after incubation on ice for a minimum of 15 mins. The precipitated protein was collected by centrifugation at 15000 rpm for 20 mins at 4°C and subsequently washed twice with 100 µL acetone. The pellets were allowed to dry in the hood for 60 mins and resuspended in 100 µL of 50 mM Tris-HCl pH 7.5, 0.1 M NaCl (**Buffer 1**). The suspension was mixed at ratio of 3:1 (v:v) in 4x LDS sample buffer (Invitrogen), heated at 95°C for 5 mins and subsequently cooled to room temperature and kept at -20°C until further use.

4.2.9.2 Preparation of total cell protein

Total cell protein sample was prepared to identify proteins that were expressed in both soluble and insoluble fraction. In this sample preparation, IB in the insoluble fraction were solubilized in the LDS sample buffer. 1 mL of bacterial culture was harvested by centrifugation at 5000 g for 20 mins at 4°C. Cell pellets were resuspended in 0.25 mL of ice-

cold Buffer 1 and sonicated for 20s at 30% amplitude, to lyse the cells. Cells were kept on ice throughout the process. Sonicated samples were mixed at ratio 3:1 (v:v) in 4x LDS sample buffer (Invitrogen), heated at 95°C for 5 mins and subsequently cooled to room temperature and kept at -20°C until further use.

4.2.9.3 Preparation of soluble protein fraction

1 mL of bacterial cultures was harvested by centrifugation at 5000 g for 20 mins at 4°C. Cell pellets were resuspended in 0.25 mL ice-cold of Buffer 1 and sonicated for 20 sec at 30% amplitude to lyse the cells. Cells were kept on ice throughout the process. Sonicated samples were spun at 13000 rpm for 15 mins at 4°C and the supernatant was collected as the soluble fraction or crude extract. The soluble fraction was mixed at ratio 3:1 (v:v) in 4x LDS sample buffer (Invitrogen), heated at 95°C for 5 minss and subsequently cooled to room temperature before kept at -20°C until further use. Crude extract for use in enzymatic assays was stored at 4°C until further use.

4.2.10 Determination of protein concentration

Protein concentration of the samples was determined using the Quick Start Bio-rad protein assay (Bio-Rad), which is based on the Bradford method (Bradford, 1976). Bovine serum albumin (BSA) solutions at concentrations from 0 to 2 mg mL^{-1} were used to prepare a standard curve.

4.2.11 SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

Protein samples were loaded into 4 % stacking and 12 % resolving SDS-polyacrylamide gel (Appendix A). BenchMark™Pre-Stained protein ladder was used as protein molecular weight standard. The gel was run in 1x SDS running buffer system at 70 V until it reached the resolving gel, when the voltage was increased to 150 V. The electrophoresis was stopped when the dye front reached the bottom of the gel.

4.2.12 Solubilization of protein that was expressed as inclusion bodies

Two approaches were used to solubilize the protein that was expressed as IB (Figure 4.1). The first approach used a range of denaturants and solubilising agent to solubilize the IB. The solubilized protein was then refolded in numbers of refolding buffers and tested for enzymatic activity. In the second approach, the same gene was expressed in a different expression vector (pET43a) that was known to have the properties to solubilize protein that previously expressed as IB. The gene was cloned with histidine tag at both C and N-terminal and referred to as Subt9195NusHisC.

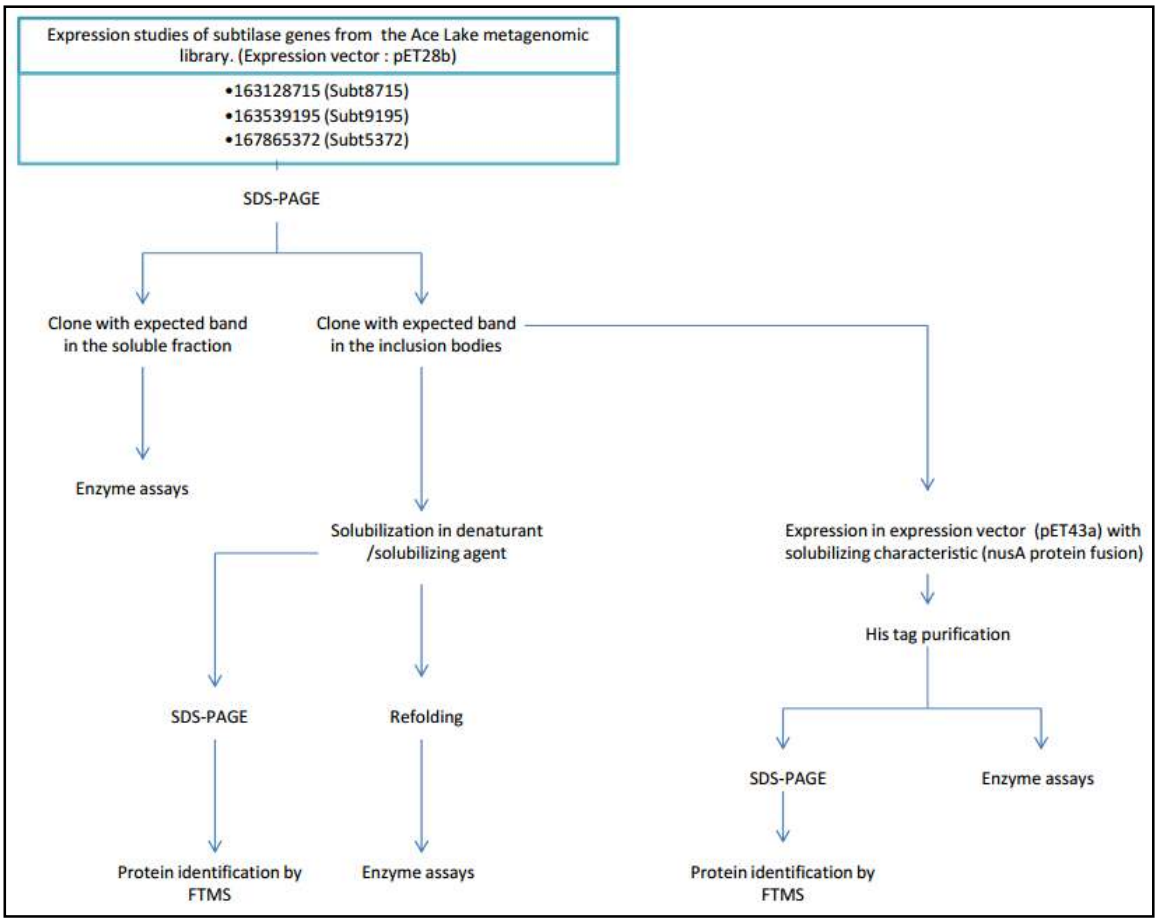


Figure 4.1: Procedures for the attempt to solubilize the protein that was expressed as IB.

4.2.12.1 Preparation of inclusion bodies for solubilization with denaturants and solubilizing agents

Cell pellets from 50 mL cultures was resuspended in 2.5 ml of ice-cold 50 mM Tris-HCl (pH 7.5)(**Buffer 2**). Cell suspensions were sonicated for 20 secs at 40% amplitude in five cycles with one mins rest between each cycle, to lyse the cells. Cells were kept on ice throughout the

process. Insoluble fractions were separated by centrifugation at 13000 rpm for 30 mins at 4°C.

4.2.12.1.1 *Recovery of solubilized protein fraction and refolding*

Pellets of insoluble fractions were resuspended in 2.5 mL of Buffer 2 containing 1% Triton X-100 by gentle mixing. The suspensions were centrifuged for 20 mins at 13000 rpm at 4°C to remove the remaining soluble proteins. This step was repeated twice. The pellets from the Triton X-100 washes were then resuspended in 2.5 ml Buffer 2 (without Triton X-100), separated into ten tubes and centrifuged at 13000 rpm for 20 mins at 4°C. The pellets from this step were then separately resuspended in Buffer 2 supplemented with denaturants (see Table 4.3), mixed gently for 5 mins and centrifuged at 13000 rpm for 20 mins at 4°C. This step was repeated twice. Pellets were then washed in the 0.25 mL of Buffer 2 (without denaturants). Finally, pellets from each tube were resuspended in Buffer 2 containing solubilising agent (Table 4.3), and gently mixed for 60 mins at 4°C. Soluble fractions were separated by centrifugation at 13000 rpm for 20 mins at 4°C. The solubilized protein was renatured or refolded *in vitro* by dilution at 4°C overnight in a range refolding buffers (Table 4.3). Proteolytic activity was measured as described in section 4.2.14.2.

Table 4.3: List of denaturants, solubilising agents and refolding buffers used in the protein solubilization and refolding process.

List of denaturants	Solubilising agent	Refolding buffers
1M Urea	1M Urea	50mM Tris-HCl (pH 7.5)
2M Urea	2M Urea	50Mm Tris-HCl (pH 7.5), 100mM NaCl
4M Urea	8M Urea	50mM Tris-HCl (pH 8.0), 100mM NaCl
0.01% sarkosyl	0.3% sarkosyl	50mM HEPES NaCl (pH 7.5)
0.05% sarkosyl		
0.1% sarkosyl		
10mM CHAPS		
20mM CHAPS		

4.2.12.2 Heterologous expression of recombinant Subt9195NusAHisC protein for His-tag purification

Overproduction of recombinant Subt9195NusHisC in *E. coli* Rosetta 2 (DE3) was performed in 2 L conical flask containing 500 mL of LB medium, 100 µg/mL of ampicilin and 34 µg/mL chloramphenicol. The culture was incubated at 37°C with shaking at 200 rpm until OD600

nm reached between 0.8-1.0 then the cultivation temperature was lowered to 25°C. After 30 mins, the cells were induced with 1 mM final concentration of IPTG overnight at 25°C. Cells were harvested by centrifugation at 5000 g for 30 mins at 4°C. In the attempt to scale up the culture volume to produce more protein, 3 L expression cultures were grown in the six 2 L conical flask containing 500 mL of LB medium, 100 µg/mL of ampicilin and 34 µg/mL chloramphenicol. .

4.2.12.2.1 *Protein sample preparation*

The cells containing protein of interest were resuspended in 25 ml of ice-cold 50mM HEPES (pH 7.5), 0.1 M NaCl **(Buffer A)**. The cell suspension was passed through the French pressure cell press (Thermo) twice to ensure thorough cell lysis. Tubes containing lysed cells were kept on ice throughout the process. Insoluble fractions were separated by centrifugation at 13000 rpm for 30 mins at 4°C. The cell lysate was filtered with through a 0.45 µm membrane before purification of the recombinant protein by immobilised metal ion chromatography (IMAC). In the scale up experiment, the culture was harvested, freeze-dried using the liquid nitrogen and kept at -80°C until further use. The purification on the affinity column was later performed on three different batches of cell lysate that were prepared using french pressure cell press.

4.2.12.2.2 *His-tag-protein fusion purification*

The purification steps were performed on a 5 ml HiTrap Chelating Column (GE Healthcare) charged with Ni^{+}, using the Akta system (Amersham Biosciences, Freiburg, Germany). The filtered cell lysate was loaded on the column that was pre-equilibrated with buffer A (1 M HEPES (pH 7.5), 0.1 M NaCl). Elution was achieved across a 25 column volume (CV) gradient 0-1M Imidazole. Eluted fractions were collected and analysed by SDS-PAGE. Fractions containing protein were dialysed and concentrated using Amicon Filtration Unit (5 kDa cut-off Millipore) in buffer A, the protein was stored at 4°C for further use.

4.2.13 Protein identification by Mass Spectrometry

NuPAGE® 4–12% Bis-Tris (invitrogen) pre-cast polyacrylamide gels were used for preparing proteins for mass spectrometry analysis. The gel was run in 1x MOPS buffer (Invitrogen) at

150 V for 30 mins. The specific protein band of interest was cut from the gel and sent to Bioanalytical Mass Spectrometry Facility (BMSF) UNSW for FTMS analysis.

4.2.14 Enzymatic assays

4.2.14.1 Skimmed milk agar assay

LB agar containing 2% skimmed milk (w/v) was spread with 100 µL of 0-2 mM IPTG Solution. The overnight culture of the recombinant subtilase were patched on the agar and incubated at 25 and 30°C for 5-7 days. Halos that formed surrounding the colony indicated proteolytic activity. *E. coli* containing pET28b without subtilase gene was used as negative control while *E. coli* containing a protease gene, pSapSH (Kulakova *et al.,* 1999) was used as positive control.

4.2.14.2 Proteolytic assay

The enzymatic assays were performed at room temperature as as described in (Tanaka *et al.,* 2008) with modifications. The reaction mixture (100 µL) contained 10 µL of cell crude extracts, 50 mM HEPES (pH 7.5) or 50 mM Tris-HCl (pH 7.5-8.8), 0.1 M NaCl and 2 mM N-succinyl-AAPF-ρ-nitroanilide (AAPF). After at least 30 mins, the amount of *p*-nitroaniline released from the substrate at room temperature was determined from the absorption at 410 nm using UV spectrophotometer (Spextra Max 340). Crude extract of the *E. coli* strain contain pET28b without subtilase gene was used as negative control. The assays were performed in triplicate.

4.2.14.3 Optimum temperature

Optimum temperature assays were performed as previously described in 4.2.14.2 using the following assay temperatures (°C): 5, 10, 20, 30, 35, 40, 45, 50, 55, 60. The reactions mixture (100 µL) contained 10 µL of enzymes, 50 mM Tris-HCl (pH 8.8), 0.1 M NaCl and 2 mM N-succinyl-AAPF-ρ-nitroanilide (AAPF).

4.2.14.4 Inhibition test of the protease using phenylmethylsulfonyl fluoride

The enzyme sample was treated with 10 µL of 100 mM phenylmethylsulfonyl fluoride (PMSF) and subjected to the assay as in section 4.2.14.2

4.3 Results

The results section includes the analysis of result from the expression studies of the three subtilase genes, Subt9195, Subt8715 and Subt5372, that were successfully cloned in pET28b vector. Following the results of expression studies, solubilization of Subt9195 that expressed as insoluble fraction is described. Temperature profile analysis was only performed on the crude Subt8715 that was expressed as soluble active cytoplasmic protease. Bioinformatic analysis pertaining to physico-chemical properties is also included at the beginning of the results section for each Subt9195, Subt8715 and Subt5372. For future work purposes, available bioinformatics data for Subt9195 and Subt8715 were utilized for protein structure prediction in the recognized public database and the results are included in the result section.

4.3.1 Analysis of 163539195 (Subt9195)

4.3.1.1 Prediction of physico-chemical properties of Subt9195

The ORF of the subtilisin-like gene, Subt9195 (1230bp) encodes a 410 aa protein with a calculated mass of 41.8 kDa, GRAVY index of 0.259 and pI of 5.19. The instability index (II) was computed to be 30.33, which classified the protein as stable. Analysis using Signal IP (Petersen *et al.,* 2011) predicted a signal peptide with a cleavage site most likely between Ala(30) and Asp(31) (Subt9195 numbering). Kyte Doolitle hydropathy plot indicated the possibility of a transmembrane region at both the C and N-terminal ends (Figure 4.2). The transmembrane region identified at the N-terminal (Figure 4.2) is the signal peptide region. It is a common problem in transmembrane protein topology that the signal peptide is predicted to be a transmembrane region because of the high similarity between the hydrophobic regions of a transmembrane helix and the signal peptide (Käll *et al.,* 2004). The transmembrane region at the C-terminal was unexpected as Subt9195 was predicted to be extracellular protease.

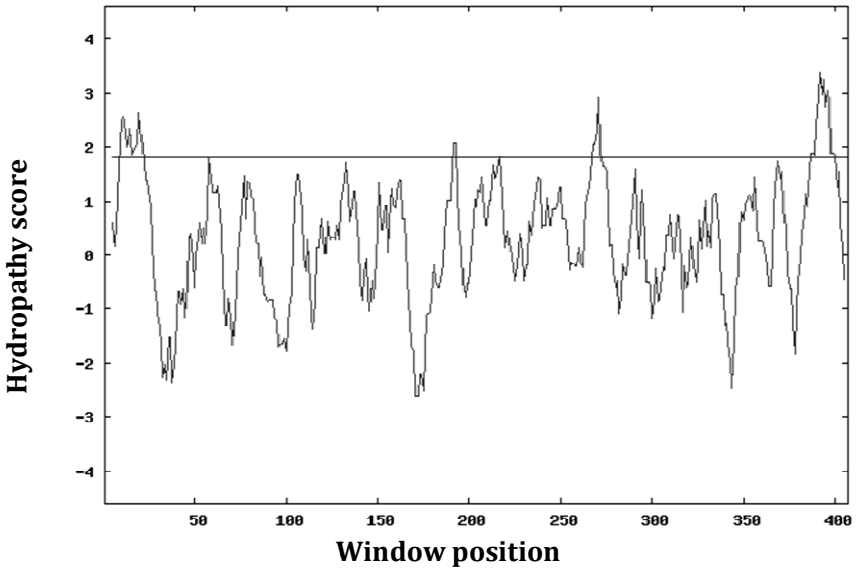


Figure 4.2: Kyte and Dolittle hydropathy plot of Subt9195 protein.

4.3.1.2 Expression studies of Subt9195

The expression studies of Subt9195 were performed at 10, 16, 25 and 30°C. The expression and solubility of the recombinant proteins were examined by SDS-PAGE. As illustrated in Figure 4.2 and 4.3, in the expression studies at 25°C and 30°C, a protein band sized ~50kb was identified in the total cell protein fraction (Figure 4.3, lane 3 and 4; Figure 4.6, lane 4-6). This expression band was not detected in the extracellular protein fraction (Figure 4.4), soluble fraction (Figure 4.5) or in any cell fraction of the uninduced culture or the control culture. The extracellular protein fraction was concentrated 50x to increase the possibility to detect low extracellular protein expression. The results indicated that recombinant Subt9195 was expressed only as insoluble protein. A similar result was obtained when the *E. coli* strain was changed from BL21 (DE3) to Rosetta 2 (DE3) (data not shown) or when the concentration of the inducer IPTG was adjusted from 0.01 mM-1 mM (Figure 4.3-4.6).

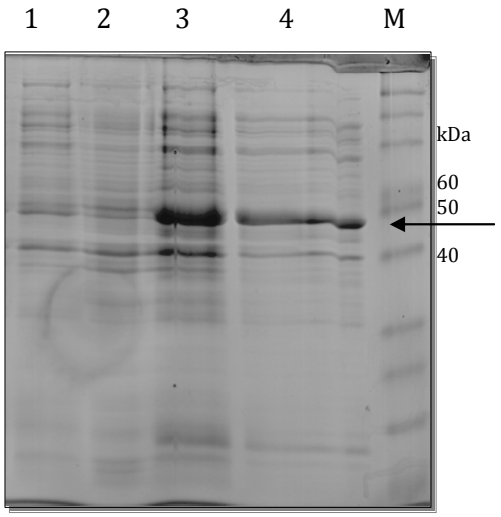


Figure 4.3: SDS-PAGE of total cell protein in expression studies of Subt9195 at 25°C in *E.coli* BL21 (DE3 1: Uninduced pET28b; 2-4: pET28b_Subt9195 (0.01, 0.1 & 1mM IPTG) M: Bench Mark™Pre-Stained protein ladder. Arrow shows the expressed protein.

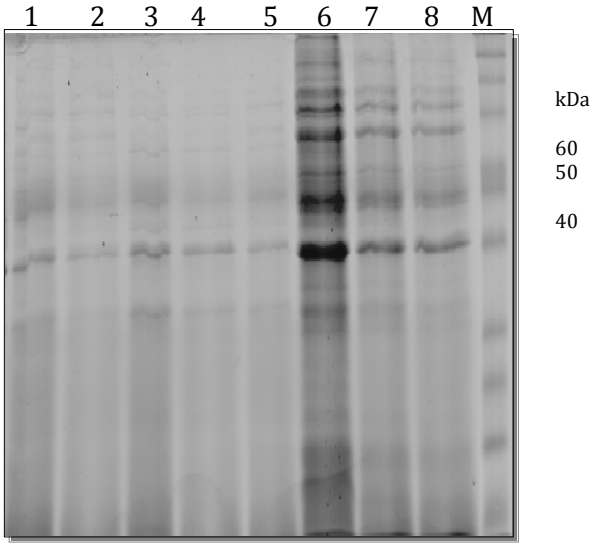


Figure 4.4**:** SDS-PAGE of the concentrated extracellular protein fraction from expression studies of Subt9195 at 25°C in *E. coli* BL21 (DE3); a) Concentrated extracellular protein: lane 1: Uninduced pET28b, lane 2-4: pET28b (0.01, 0.1 & 1mM IPTG), lane 5: Uninduced pET28b_Subt9195, lane 6-8: pET28b_Subt9195 (0.01, 0.1, 1mM IPTG).

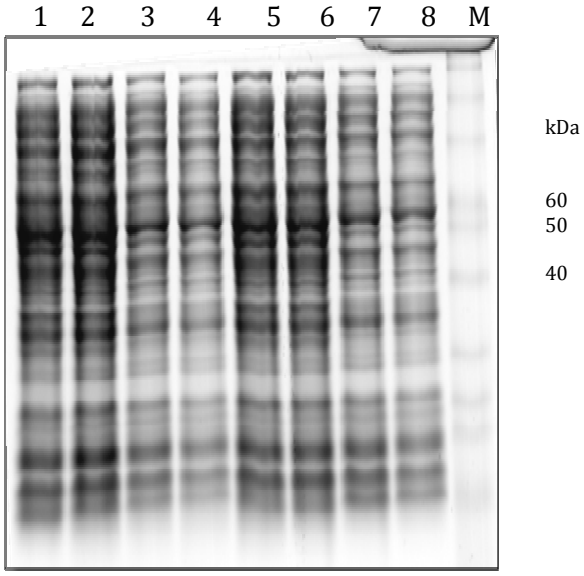


Figure 4.5**:** SDS-PAGE of the soluble fraction from expression studies of Subt9195 at 25°C in *E. coli* BL21 (DE3); lane 1: Uninduced pET28b_Subt9195, lane 2-4: pET28b_Subt9195 (0.01, 0.1, 1mM IPTG); lane 5: Uninduced pET28b , lane 6-8 : pET28b (0.01, 0.1, 1mM IPTG) M: Bench Mark™Pre-Stained protein ladder.

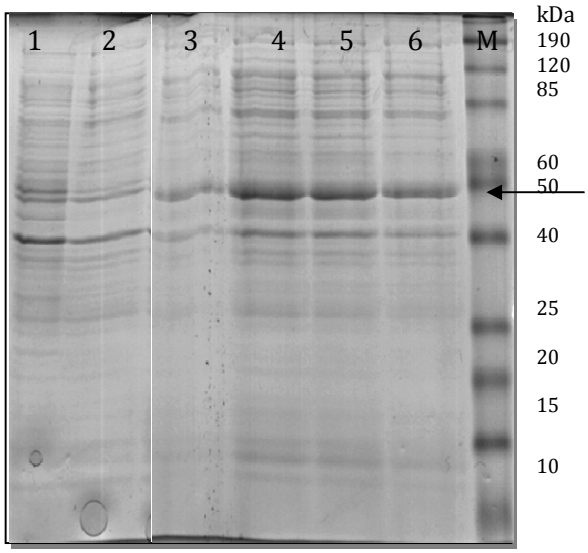


Figure 4.6**:** SDS-PAGE of total cell protein from expression studies of Subt9195 at 30°C in *E. coli* BL21 (DE3). Lane 1: Uninduced pET28b, lane 2: pET28b (0.01mM IPTG), lane 3: Uninduced pET28b_Subt9195, lane 4-6: pET28b_Subt9195 (0.01, 0.1 and 1 mM IPTG), M: BenchMark ™Pre-Stained protein ladder. Arrow showed the expressed protein.

In the attempt to obtain active and soluble Subt9195 protein, the cultivation temperature was dropped to 10°C. The expression profile of the clone Subt9195 at 10°C was recorded for 220 hours after induction with 0.5 mM IPTG and compared to the uninduced culture. Figure 4.7 illustrates the slow growth increase of the culture containing Subt9195 clone and growth decline at 200 hours after induction which showed OD of 1.38 at 600 nm. After 40 hrs, the uninduced pET28b and pET28b + Subt9195 reached OD=2. pET28b induced with 0.5 mM IPTG reached OD=2 after 120 hours. Analysis of protein expression by SDS-PAGE indicated that Subt9195 was not expressed at 10°C. The expression band was not detected either in the soluble fraction, the total cell protein or the concentrated extracellular protein (Figure 4.8).

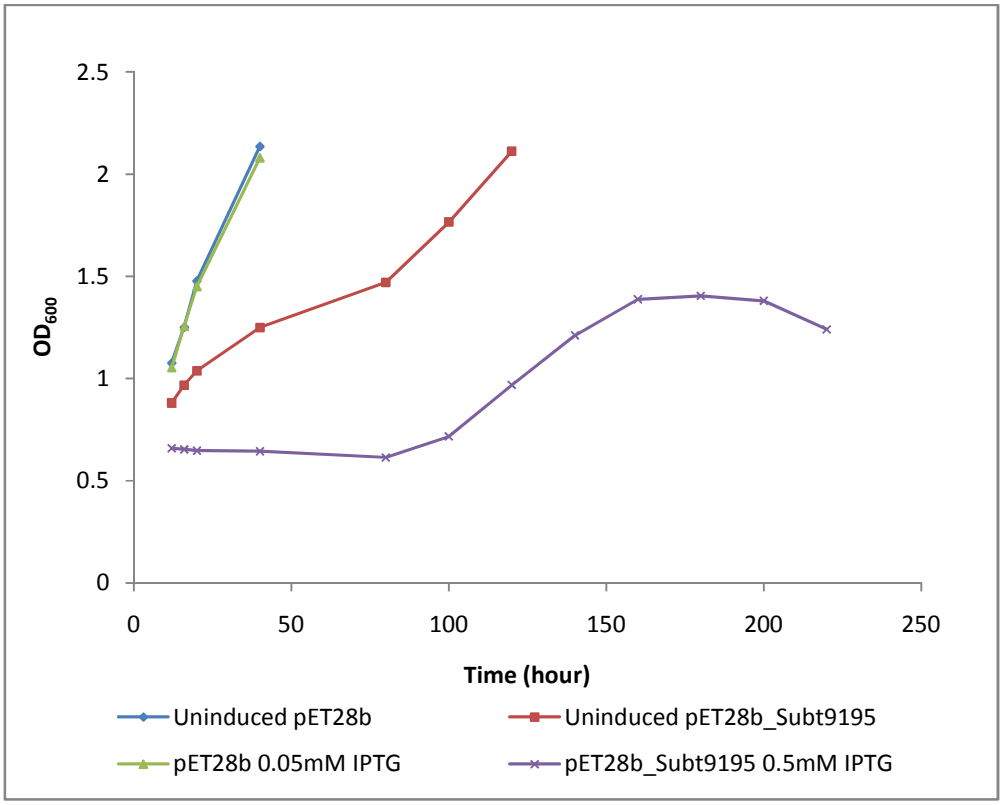


Figure 4.7: Growth profiles of pET28b_Subt9195 in the expression studies at 10°C.

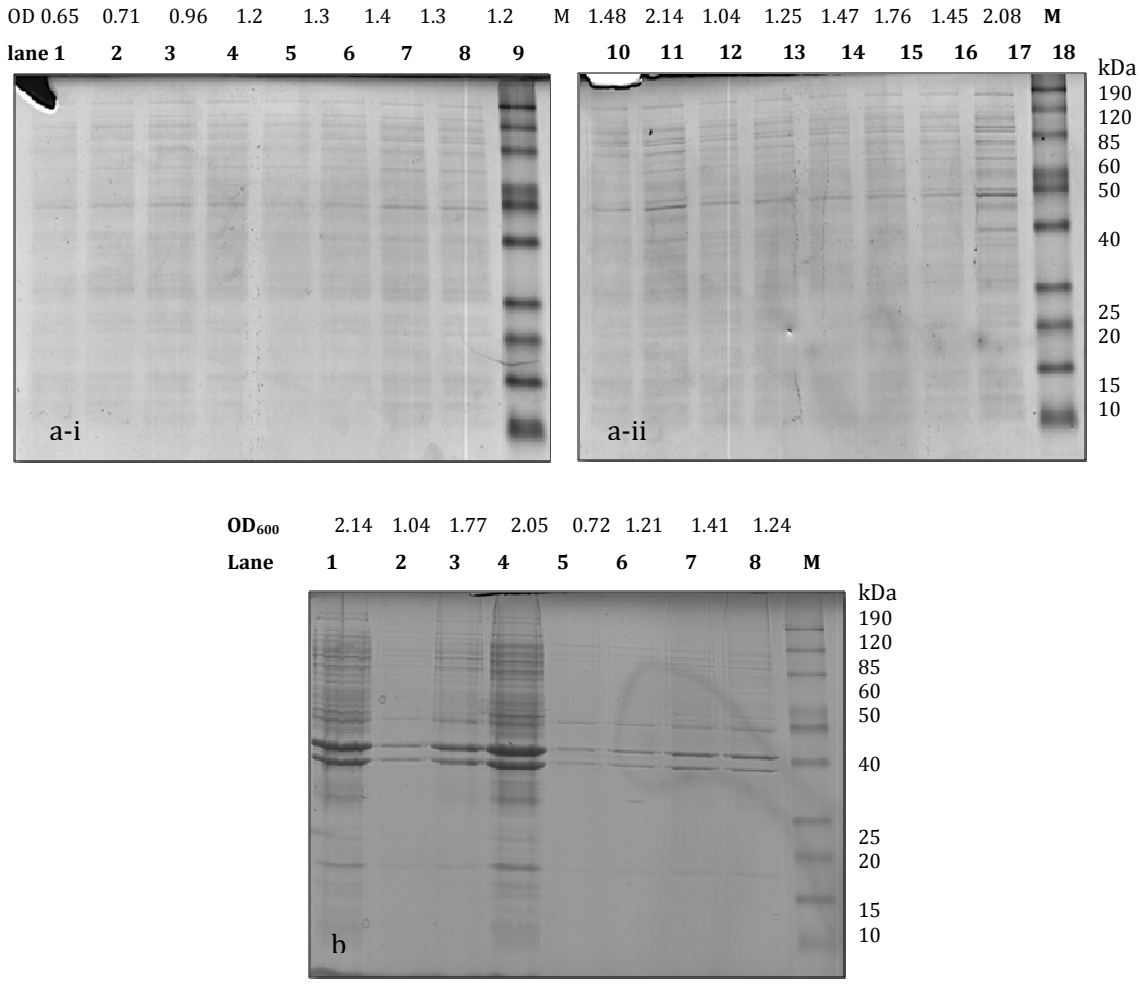


Figure 4.8: SDS-PAGE of expression studies of Subt9195 at 10°C as detected by taking OD readings at 600 nm. a-i) Soluble fraction; lane 1-8: pET28b_Sub9195 (0.5 mM IPTG), a-ii) Soluble fraction; lane 10-11; Uninduced pET28b, lane 12-15: pET28b (0.5 mM IPTG), lane 16-17: Uninduced pET28b_Subt9195. b) Total cell protein; lane 1: Uninduced pET28b, lane 2-3: pET28b (0.5 mM IPTG), lane 4: Uninduced pET28b_Subt9195, lane 5-8: pET28b_Subt9195 (0.5 mM IPTG), M: BenchMark ™Pre-Stained protein ladder.

4.3.1.3 Skimmed milk agar assay

In order to detect the expression of protease activity, cultures of Subt9195 were grown on LB skimmed milk agar containing IPTG at concentrations ranging from 0-2mM. The cultures were incubated at 25 and 30°C to investigate the effect of temperature on the protein expression. As illustrated in Figure 4.9a, at 25°C, colonies containing Subt9195 clone showed halo formation around the colony on the plate containing 1 mM IPTG (Figure 4.9a (d)) while

at 30°C, halo formation was observed around the culture on plate containing 0-1.5 mM IPTG after 5 days of incubation (Figure 4.9b (a-d). The formation of halos surrounding Sub9195 culture indicated the expression of active protease. The halo was not identified around pET28b culture.

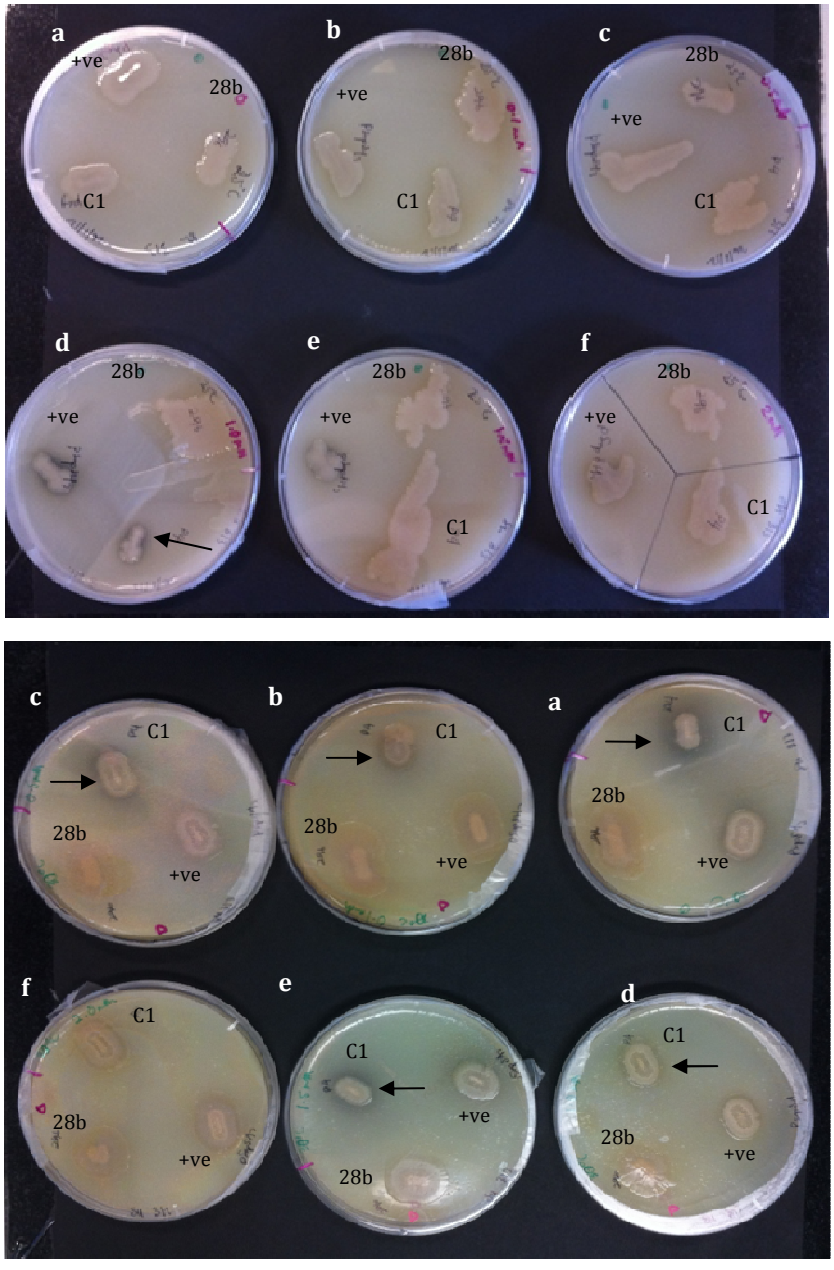


Figure 4.9: Skimmed milk agar assay of culture Subt9195 at 25 and 30°C. i: 25°C, ii: 30°C; a-f: IPTG concentration (mM); a: 0, b: 0.1, c: 0.5, d: 1.0, e: 1.5, f: 2.0. 28b=Culture of pET28b; C1= Culture of pET28b_Subt9195, +ve= Positive control pSapSH.

4.3.1.4 Solubilization of Subt9195 protein in urea and Sarkosyl

As shown previously in section 4.3.1.2, Subt9195 was expressed only in the insoluble cell fraction. However, halo formation in the skimmed milk agar assay indicated that Subt9195 might express active protease. In order to obtain soluble Subt9195 protein, the IB from the expression of Subt9195 at 25°C was subjected to solubilization in 0.3% sarkosyl, 0.1M, 2M, and 8M urea. The effect of the prewashing step on the success of solubilization using Triton X-100, urea, sodium lauroyl sarcosinate (Sarkosyl) and 3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate (CHAPS) was apparent. As illustrated in Figure 4.10, after solubilization with 8 M urea, a ~50 kDa sized band was identified in the soluble fraction. The size of the protein band was a similar size to the predicted Subt9195 (~41.8 kDA) reported in section 4.3.1.1 Based on the intensities of the protein band following the SDS-PAGE, the sample that was prewashed with 1 M urea showed the highest solubility (lane 1, Figure 4.10).

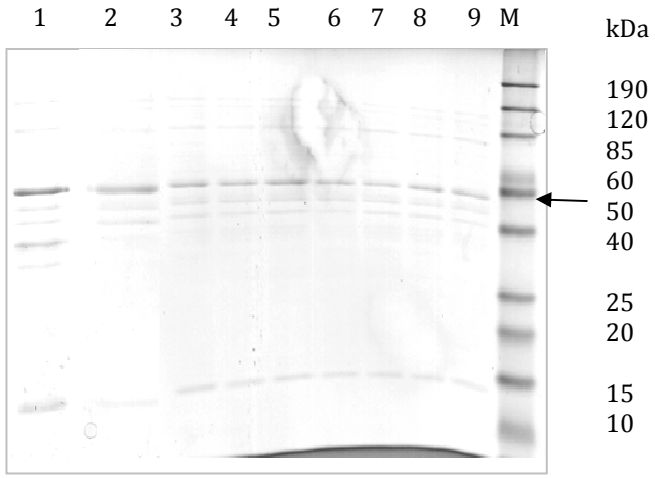


Figure 4.10: SDS-PAGE of solubilization of Subt9195 protein in 8M urea. The numbering represents the detergents and denaturants being used before solubilization with 8M Urea. 1: 1M urea 2: Nil denaturant 3: 2M Urea 4: 4M Urea 5: 0.01% Sarkosyl 6: 0.05% sarkosyl 7: 0.1% Sarkosyl 8: 10mM CHAPS 9:20mM CHAPS M: BenchMark™Pre-Stained protein ladder. The soluble protein is indicated by an arrow.

In a further attempt to solubilize the protein in a milder solubilising agent (1 M urea, 2 M urea and 0.3 % sarkosyl), 0.3 % sarkosyl successfully solubilized the protein regardless of the prewashing step using 1 M Urea or 0.05 % sarkosyl. As shown in Figure 4.11, a protein ~50 kDa band was identified in each of the lanes marked 3 where 0.3 % sarkosyl was used as the solubilising agent. No band was detected in any of the lanes marked 1 or 2 where 1 M and 2 M urea were used as the solubilising agent. The protein bands with size ~50 kDa were cut from the gel and sent for FTMS analysis. The identity of the protein was confirmed as soluble recombinant Subt9195. The analysis identified multiple peptides that matched with high individual peptide scores as well as an excellent score for the whole Subt9195 deduced protein sequence. Figure 4.12 summarised the identified peptides fragments that showed matches to Subt9195 deduce protein sequence.

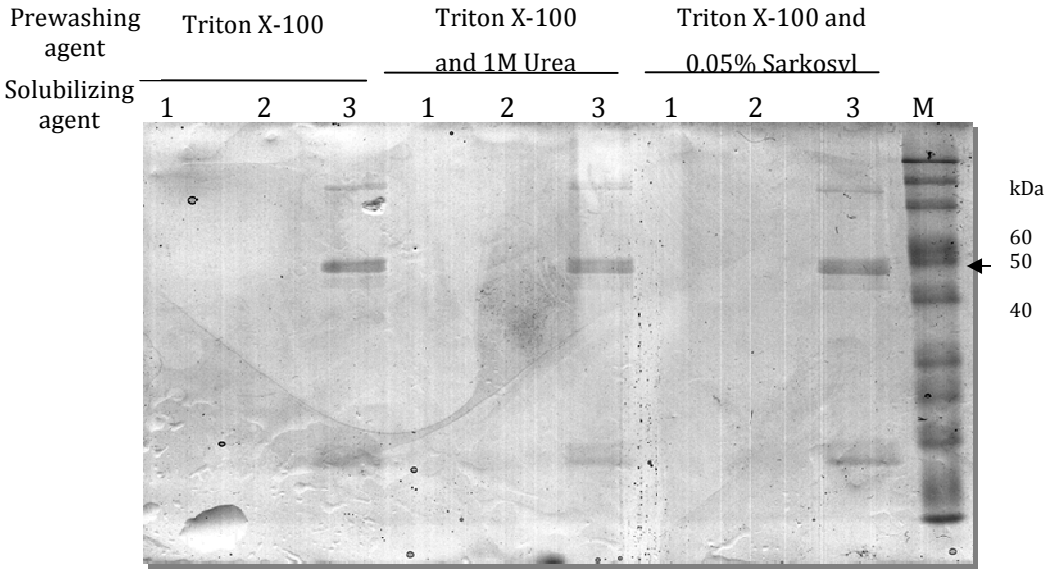


Figure 4.11**:** SDS-PAGE of solubilization of IB of Subt9195 following prewashed with Triton X-100, 1M urea and 0.05% sarkosyl. The numbering represents the solubilising agent; 1: 1M Urea, 2: 2M Urea and 3: 0.3% sarkosyl. M: BenchMark™Pre-Stained protein ladder. The soluble protein is indicated by an arrow.

MLKKNPILVWFLALSLGIAPVFVVSSPAQADQVRDRQYWLQDYGIEQAWAITRGAG
VRIAIIDTGVDGSHQDLEGAVVAGADFSGLGSTNGQTPVGSDRRHGTMVASLAAGR
GNGVQNGVIGSAPEAEIISASISFGGGAVSPDDQIARAVRFAVDAGADVISLSLTRNT
RDWPETWDDAFTYAADRDVVVIAAAGNRGSGTVAVGAPATMPGVLAVGGVTQEG
IASDAASSQGISLGVMAPSEGLVGAIPGGGYVSWSGTSGAAPIVAGIAALVRAAYPRM
SADNVINRILVSARPVSDQVPDPLYGYGLVNAYEALTREVPSVSANPLGALDAWITL

Figure 4.12: Peptides fragment (underlined) that matched to the deduced amino acid sequence of Subt9195 in the FTMS analysis.

4.3.1.5 *In vitro* Protein Refolding

The solubilized protein in 8 M urea was diluted 30x, while solubilized protein in 0.3 % sarkosyl was diluted to yield 0.01 % sarkosyl solution with refolding buffers to refold the protein into its native state. The success of *in vitro* refolding process was determined by identifying protease activity in the protease assay described in section 4.2.14.2. Assays conducted with the refolded protein using AAPF as a substrate did not show protease activity.

4.3.1.6 Solubilization of Subt9195 using NusA-fusion protein system

In the attempt to solubilize Subt9195 protein, the gene was cloned into pET43.1a as a NusA protein fusion containing a histidine tag at both the N and C-terminal. The recombinant construct was designated as pET43aSubt9195NusAHisC. The expression study of this new construct was carried out using *E. coli* Rosetta 2 (DE3) strain, which was cultured at 37°C, induced with 1 mM IPTG at 30°C and incubated for 7 hours at 30°C with shaking at 200 rpm. The expression of recombinant protein was subsequently analysed by SDS-PAGE.

As illustrated in Figure 4.13, three bands were indentified in the soluble fraction of the induced culture. The bands sized ~100, ~ 70 and ~ 50 kDa were not found in the uninduced culture (Figure 4.13, lane 1) or the concentrated supernatant of the induced culture. In order to make certain whether the bands were the desired protein, cultures of pET43.1a were induced with 1 mM IPTG to see the expression of NusA without any fusion protein. Following SDS-PAGE, as showed in Figure 4.14, a band sized ~70 kDa was detected

in the soluble fraction of pET43.1a induced with 1mM IPTG thus confirming the band sized ~70 kDa in Figure 4.13 was NusA protein. This observation also indicated Subt9195 was expressed as soluble NusA protein fusion sized ~120 kDa. However, the presence of another two bands suggested intracellular processing of Subt9195 from the fusion protein releasing peptides sized ~50 kDa (labelled X in Figure 4.13) and ~70 kDa (labelled NusA protein in Figure 4.13). The soluble fraction was subjected to IMAC purification to isolate the Subt9195 protein.

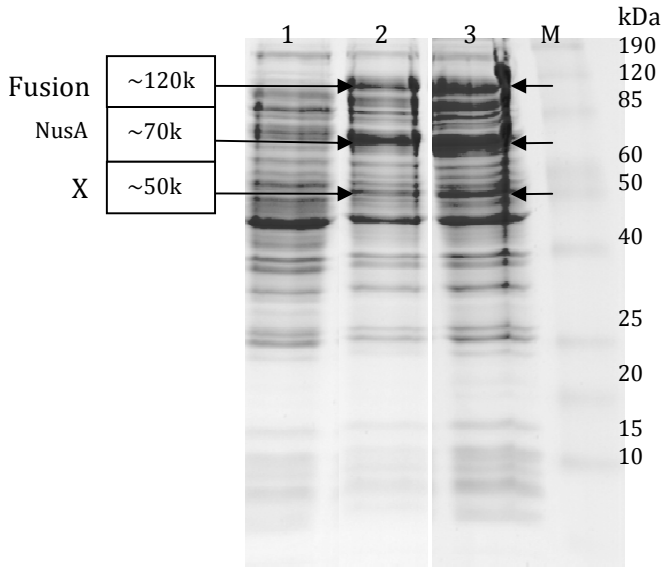


Figure 4.13: SDS-PAGE of soluble fraction of pET43aSubt9195NusHisC in the expression studies at 30°C. Lane 1: uninduced pET43a_Subt9195NusAHisC, lane 2: pET43a_Subt9195NusAHisC induced with 1 mM IPTG and lane 3: replicate of 2, M BenchMark™Pre-Stained protein ladder, X: Protein of interest.

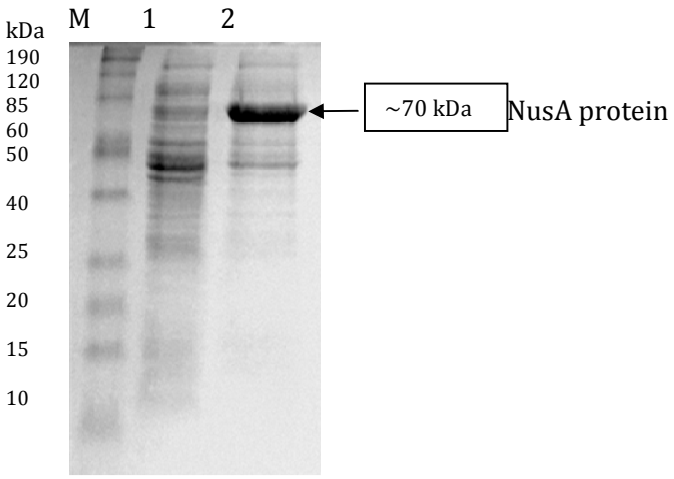


Figure 4.14**:** SDS-PAGE of soluble fraction of pET43a.1 culture inducted with 1mM IPTG at 30°C; Lane 1: uninduced pet43.1a, lane 2: pet43.1a induced with 1mM IPTG. M: BenchMark™Pre-Stained protein ladder

4.3.1.7 Protein Purification

4.3.1.7.1 *His-tag purification of the Subt9195NusAHisC recombinant protein*

Subt9195NusAHisC was designed to have a His-tag at both the N and C-terminal, thus allowing protein purification by IMAC. Taken into consideration was the possibility of signal peptide processing which would remove the N-terminal His-tag. As illustrated in Figure 4.15, the His-tag purification yield was not pure. In addition there were discrepancies in terms of the size of the expression band from Figure 4.13. SDS-PAGE analysis of the purification showed the most prominent band sized ~80 kDa in fraction 4 (Figure 4.15, lane 6). The bands sized 120, 80 and 60 kDa were observed in fraction 3 (Figure 4.15, lane 5). In fraction 4 and 5 (Figure 4.15, lane 6 and 7) there was another ~50 kDa band. Smaller molecular weight protein, ~30 kDa in size, was observed in fraction 3, 4 and 5 (Figure 4.15, lane 6, 7 and 8).

Fraction 4 from the column elution, as well as the dialysed and concentrated sample of fraction 2-6 from the His-tag purification were analysed on 4-12% Bis–Tris NuPage gels (Invitrogen). This was to achieve higher resolution for band excision and protein identification via mass spectrometry analysis as well as to detect any changes to the band pattern from the previous preparation. As shown in Figure 4.16, two discrete bands at ~90 and ~70 kDa were observed in the serial dilution of sample 4 from the column elution (lane

1-4). This protein was previously observed as the prominent ~80 kDa band (Figure 4.15, lane 6). Mass spectrometry analysis indicated that both bands were in fact NusA protein. Interestingly, the gel showed that, the prominent band was no longer visible after dialysis and concentration (Figure 4.16, lane 8). In fact, a new band sized ~50 kDa, shown to be NusA protein by mass spectrometry analysis, was evident. In addition, lower molecular weight proteins ~15 to ~45 kDa in size were identified by mass spectrometry analysis as Subt9195, which suggested that a proteolytic processing event had occurred.

The purified Subt9195NusAHisC also seemed to be self digesting when stored at 4°C. Figure 4.17 shows the SDS-PAGE analysis of fraction 4 (dialysed and concentrated fraction) after 2 weeks storage at 4°C. Previous higher molecular weight proteins were no longer evident as prominent bands, replaced by lower molecular weight proteins ~17 kDa in size, thus confirming the presence of protease activity.

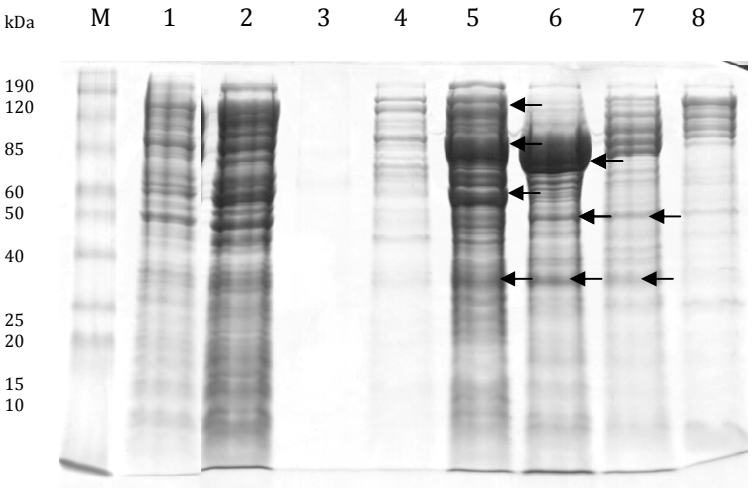


Figure 4.15: SDS-PAGE of fractions obtained after pET43a_Subt9195NusAHisC His-tag purification 1: total cell protein of crude cell extract, 2: flow through 3: Fraction 1, 4: fraction 2, 5: Fraction 3, 6: Fraction 4, 7: Fraction 5, 8: Fraction 6. M: Bench Mark™Pre-Stained protein ladder. Arrow showed the protein mentioned in the result section

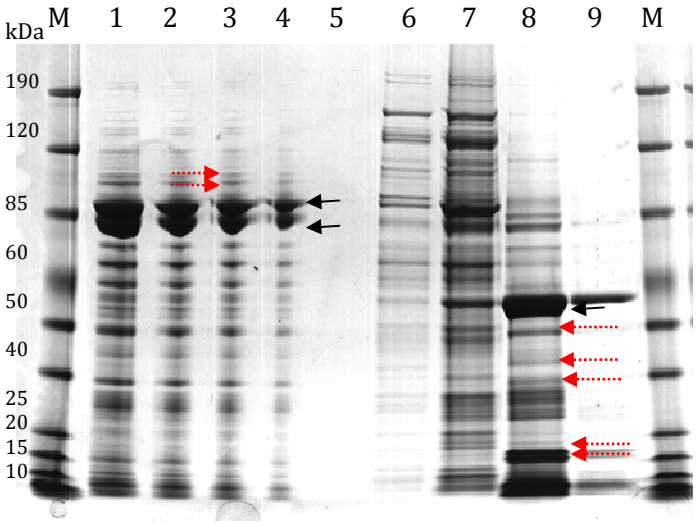


Figure 4.16: Purified fractions of pET43a_Subt9195NusAHisC before and after concentration and dialysis; on 4-12% Bis–Tris NuPage (Invitrogen) gel. M: BenchMark ™Pre-Stained protein ladder, 1-4: Serial dilution of fraction 4 before concentration and dialysis. 5-9: Fraction 2-6 after dialysis and concentration respectively. Solid arrow indicates band of NusA protein. Dashed arrow indicates band with Subt9195 protein ID.

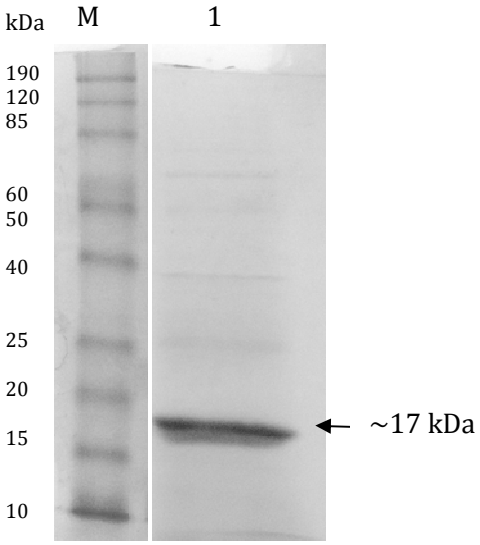


Figure 4.17: SDS-PAGE of fraction 4 after 2 weeks storage at 4°C indicate self-digesting activities.

4.3.1.8 Protease assay of partially purified Subt9195

SDS-PAGE analysis of the dialysed and concentrated Subt9195 showed a difference in the band pattern compared to undialysed and unconcentrated sample indicating the presence of protease activity in the fraction 4 (Figure 4.16, lane 6). Three protein fractions of Subt9195NusAHisC, that were partially purified as described in section 4.3.1.7, were subjected to protease assay using AAPF as the substrate. The dialysed and concentrated fractions 3, 4, and 5 (Figure 4.15, lane 6,7 and 8) were checked for protease activity. The effect of calcium ions on enzyme activity was also tested by supplementing the assay with 100 mM $CaCl_2$. The assay detected proteolytic activity in fraction 4 (Figure 4.18a). The presence of 100 mM $CaCl_2$ in the assay did not increase the enzymatic activity in any fraction (Figure 4.18b). The activity was inhibited completely with 10 mM PMSF which indicated a common characteristic of serine protease.

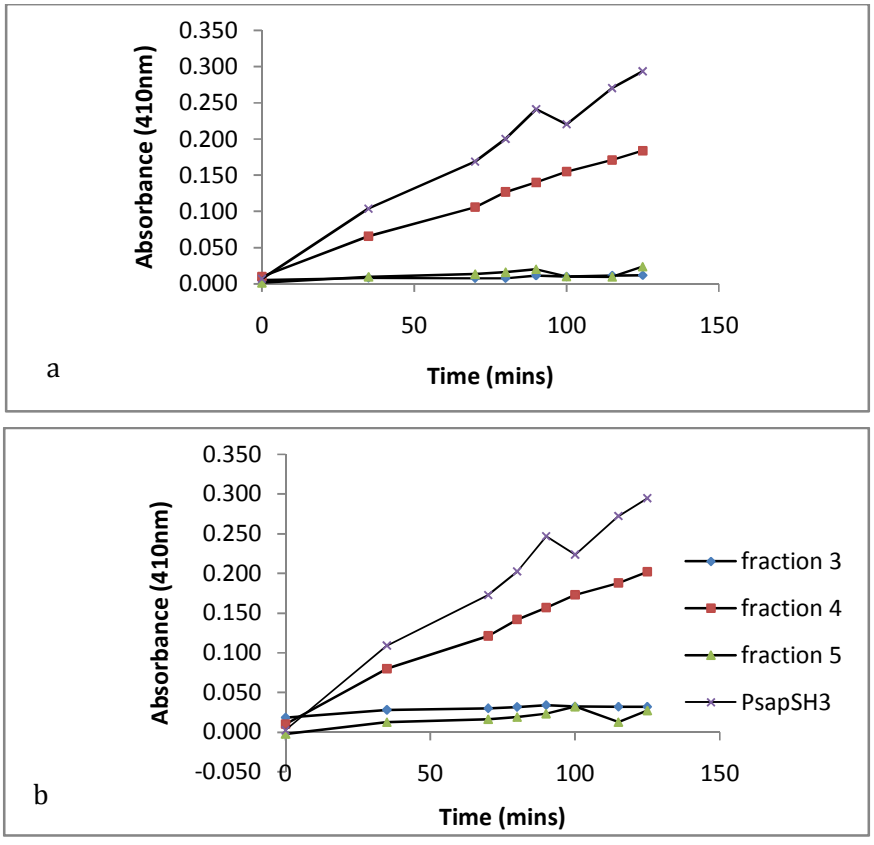


Figure 4.18: Protease activity of His-tag purified Subt9195NusAHisC at room temperature. Assay was conducted using a) 50mM HEPES (pH 7.5), 0.1M NaCl; b) 50mM HEPES (pH 7.5), 0.1M NaCl, 10mM $CaCl_2$ at room temperature.

4.3.2 Analysis of 163128715 (Subt8715)

4.3.2.1 Prediction of physico-chemical properties of Subt8715

The ORF of the Subt8715 gene (3171bp) encodes a protein comprising 1057 amino acids with a calculated mass of 105.2 kDa, GRAVY index of 0.16 and pI of 4.94. The instability index (II) was computed to be 22.18, which classified that the protein as stable. Kyte Doolitle hydropathy plot did not indicate any transmembrane region in the protein (Figure 4.18). Analysis with Signal IP (Petersen *et al.,* 2011) did not predict any signal peptide cleavage site.

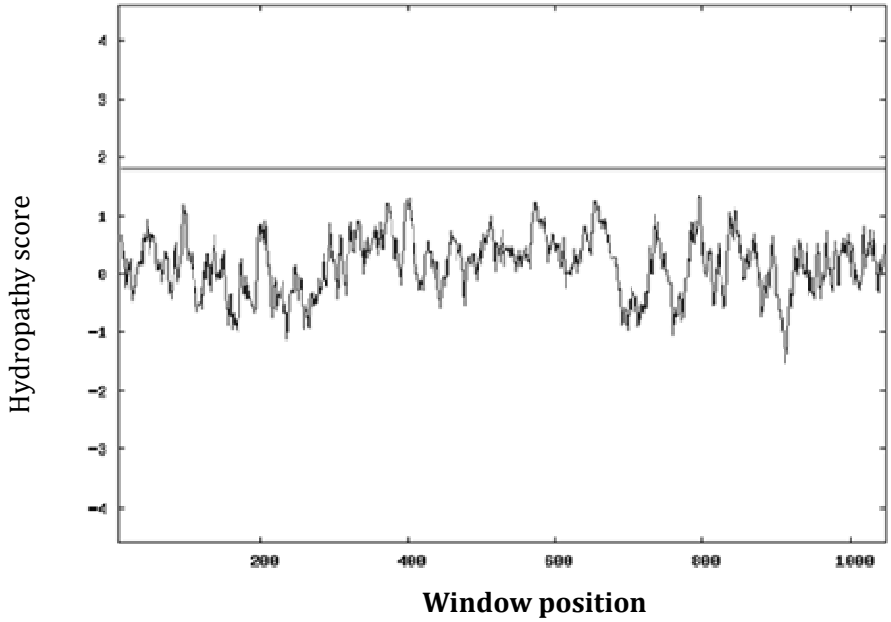


Figure 4.19: Kyte and Dolittle hydropathy plot of Subt8715.

4.3.2.2 Expression studies of Subt8715

Expression studies of the recombinant Subt8715 at 25°C and 30°C in *E. coli* BL21 (DE3) showed the expression of recombinant protein in the soluble fraction. As illustrated in Figure 4.20, a prominent ~155 kDa band was detected in the soluble fraction of the pET28bSubt8715 induced with 0.05, 0.1, 0.5 or 1 mM IPTG. The protein size was about ~50 kDa larger than the calculated molecular weight. A similar band was not detected in the concentrated extracellular protein fraction (Figure 4.21) suggesting the protein was not

expressed as extracellular protein. The recombinant protein was subjected to protease assay to detect activity.

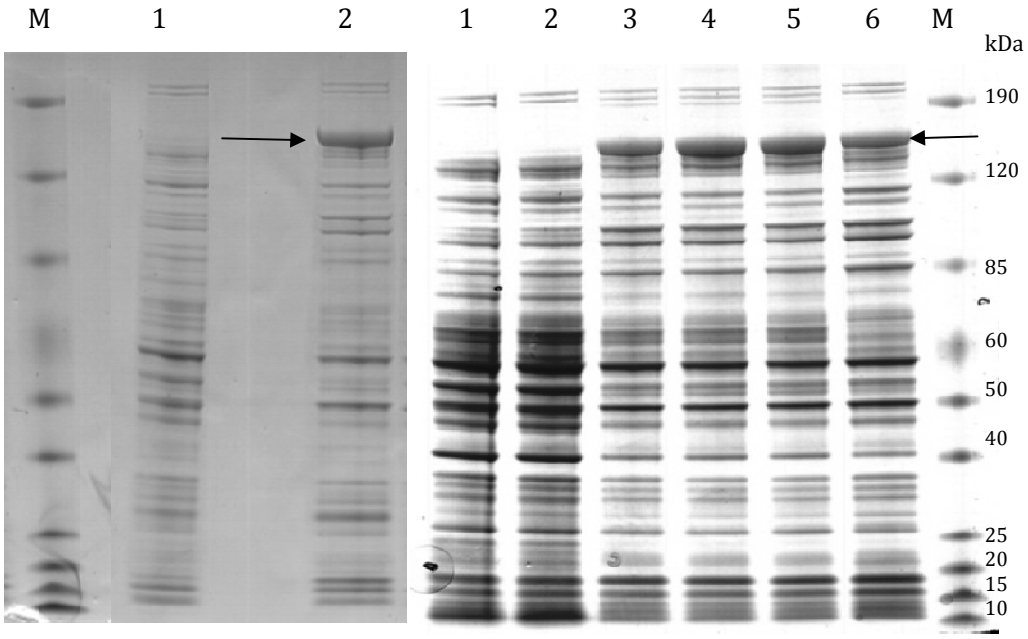


Figure 4.20: SDS-PAGE of soluble fraction of pET28bSubt8715 in BL21 (DE3) at 25 and 30°C. a) Expression at 25°C; lane 1: Uninduced pet28b, lane 2: pET28b + Subt8715 (1mM IPTG); b) Expression at 30°C; lane 1: uninduced pET28b, lane 2: uninduced pET28b + Subt8715, lane 3-6: pet28b + Subt8715 (0.05, 0.1, 0.5 & 1mM IPTG), M: BenchMark™Pre-Stained protein ladder. The arrows indicate the expression band.

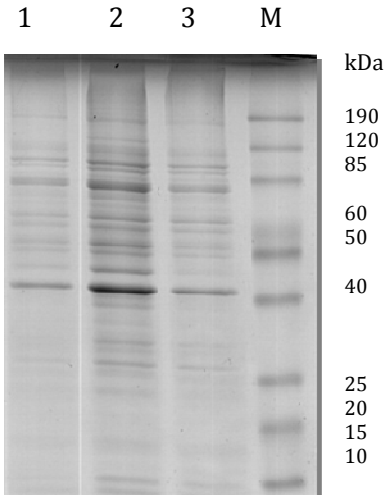
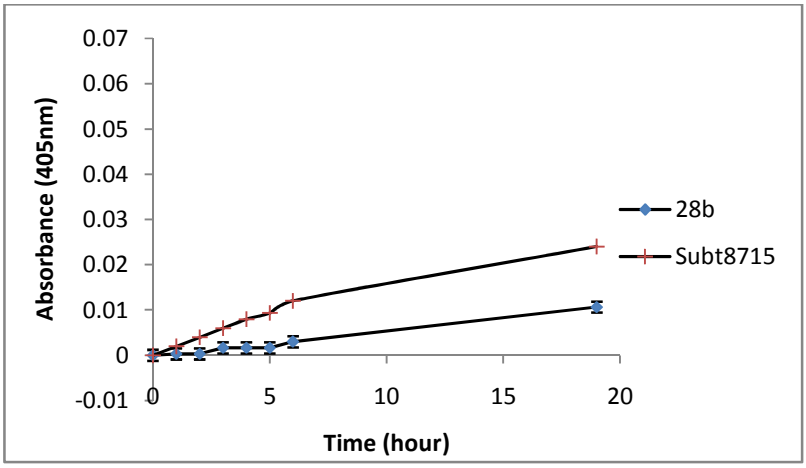


Figure 4.21**:** SDS-PAGE of extracellular protein fraction of pET28b + Subt8715 in *E. coli* BL21 (DE3). 1: Uninduced pET28b, 2: pET28b + Subt8715 (0.1mM IPTG) 3: pET28b +Subt8715 (1mM IPTG), M: BenchMark™Pre-Stained protein ladder.

4.3.2.3 AAPF assay of the cell crude extract

The enzyme activity of the recombinant subtilase on AAPF, was determined by measuring the the release of ρ-nitroaniline by detecting the increase of absorbance reading at 410 nm by 0.01. As illustrated in Figure 4.22, the crude extract of Subt8715 showed an increase in absorption at 410nm with time, indicating the presence of protease activity compared to the control (pET28b). The absorption at 410 nm was higher in the assay at pH 8.8 (Figure 4.22b) compared to pH 7.5 (Figure 4.22a). The enzymatic activity in both assays was linear for up to 6 hours.

a)



b)

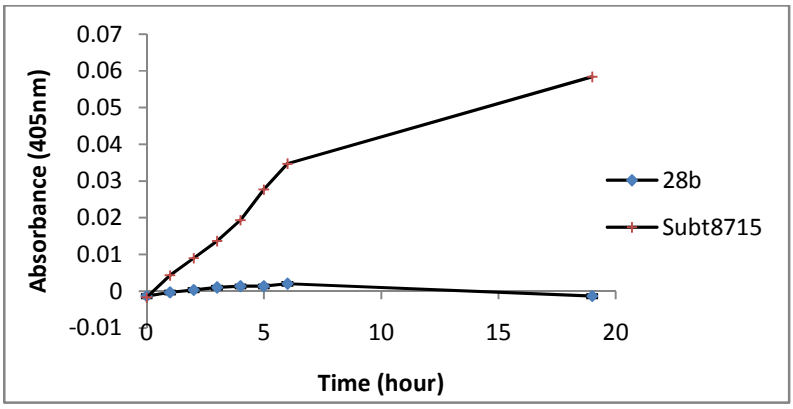


Figure 4.22: Activity assay of crude cell extracts of culture containing Subt8715 compared to the negative control (pET28b). Assays were performed at room temperature using AAPF as substrate.

4.3.2.4 Effect of temperature to Subt8715 protease activity

The effect of temperature (5-60°C) on enzymatic activity of Subt8715 was assayed to determine whether Subt8715 demonstrated characteristics of cold-adapted proteases. As showed in Figure 4.23, the optimum temperature for crude extracts of Subt8715 was 50°C. It showed increased of activity towards AAPF with temperature between 5 and 50°C and decreased sharply at 55°C. Subt8715 showed 20% and 40% residual activity at 20°C and 60°C respectively. The activation energy of the reaction catalysed by Subt8715 was determined from an Arrhenius plot of the values shown in Figure 4.24. From the slope of the graph, the calculated activation energy (E_a) of Subt8715 is 36.7 kJ/mol.

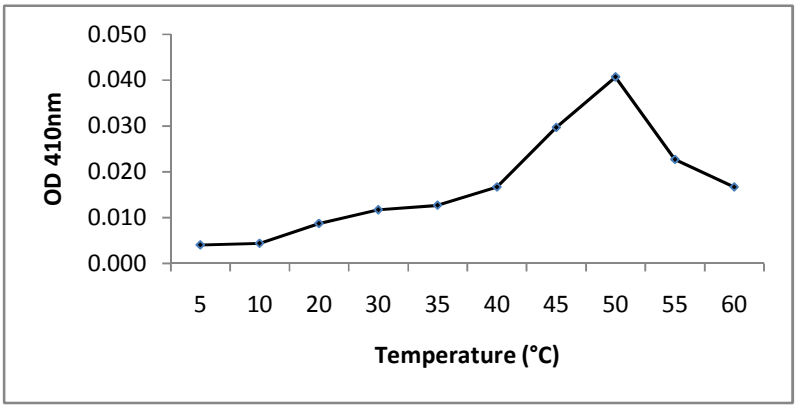


Figure 4.23: Effect of the temperature on Subt8715 activities towards AAPF.

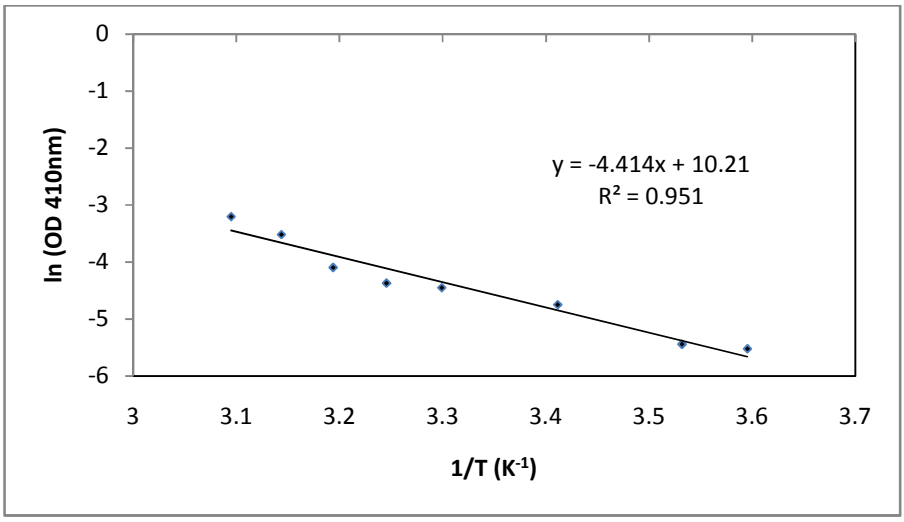


Figure 4.24: Ln OD vs 1/T of Subt8715

4.3.3 Analysis of 167865372 (Subt5372)

4.3.3.1 Prediction Physico-chemical analysis of Subt5372

Subt5372 was an ORF of 1194 bp comprising 398 amino acids. The calculated mass of this protein is 40.9 kDa while the GRAVY index value is 0.161. This protein shares 70 % sequence identity with Subt9195. Kyte Doolitle hydropathy plot did not indicate any transmembrane regions in the protein (Figure 4.25). Analysis with Signal IP (Petersen *et al.,* 2011) did not predict any signal peptide cleavage site.

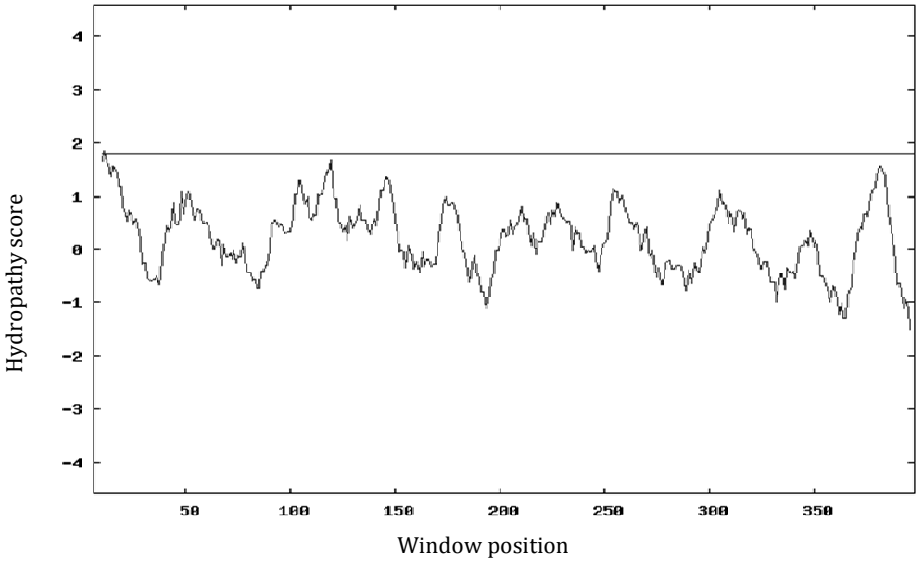


Figure 4.25**:** Kyte and Dolittle hydropathy plot of Subt8715.

4.3.3.2 Expression studies of Subt5372

Expression studies of the recombinant Subt5372 were performed using two *E. coli* strains; BL21 (DE3) and Rosetta 2 (DE3) at 25 and 30°C. Similar protein expression patterns were found in both *E. coli* strains. As illustrated in Figure 4.26a and Figure 4.26b, the ~50 kDa band was detected in the total cell protein fraction of clones induced with 0.1 mM IPTG (lane T2) and 1mM IPTG (lane T3). The band was not detected in the soluble fraction (lane S2 and S3 Figure 4.26 (a) and (b)) which indicates that the recombinant subt5372 was expressed as insoluble protein.

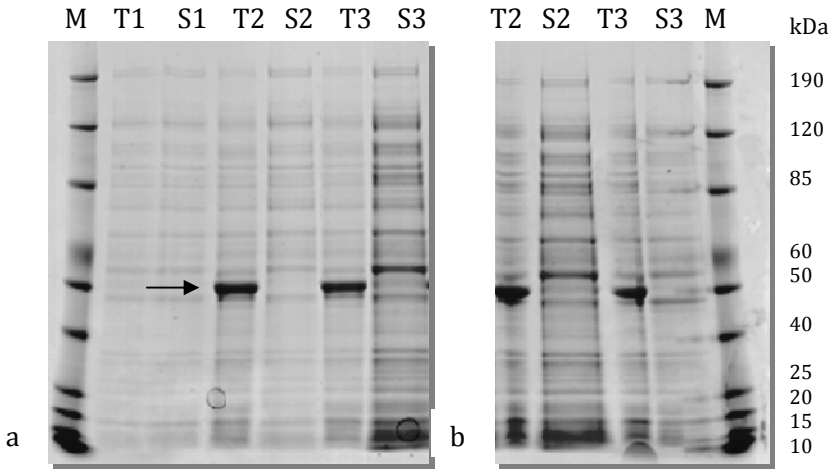


Figure 4.26: SDS-PAGE of total cell protein and soluble fraction of pet28bSubt5372 in Rosetta 2 (DE3) at a)25 and b)30°C., T: Total cell protein S: Soluble cell fraction, 1: Uninduced pet28b, 2: pET28bSubt5372 (0.1mM IPTG) 3: pET28bSubt5372 (1mM IPTG), M: BenchMark™Pre-Stained protein ladder. Arrow showed the expressed protein.

4.3.3.3 Predictions of functional and structural domains

Since, recombinant Subt9195 and Subt8715 was not yet available for purification, three dimensional modelling of these proteins were performed in I-TASSER to predict their protein structure. The confidence of the predicted structures was based on the C-score which is typically in the range of [-5, 2]. The C-score of higher value signifies a model with a high confidence and vice-versa.

4.3.3.4 Structure modelling of Subt9195

Five models of Subt9195 were computationally generated using the I-TASSER algorithm with C-scores ranging from -0.38 to -2.16, with higher scores representing higher confidence model. Model 1 (Figure 4.27) with the highest C-score, -0.38 was used for further analysis. A TM score >0.5 indicates a model of correct topology. The cluster density is defined as the number of structure decoys at unit of space in the SPICKER cluster. The predicted model has a high number of decoys and cluster density (Table 4.7), which means the structure occurs more often in the simulation trajectory and therefore signifies a better quality model.

Table 4.27 **:** Estimated accuracy of the structure model.

C-score	Exp TM Score	Exp RMSD	No. Of Decoys	Cluster Density
-0.38	0.66+-0.13	7.7+-4.3	1200	0.2030

According to the TM score, the highest structural similarity is with Tk-SP, a hyperthermostable subtilisin-like serine protease from *Thermococcus kodakaraensis* (3AFGA). This protein has 35% sequence identity to Subt9195. The TM score of 0.931, which is > 0.5 indicated the structures share the same fold. Figure 4.28 shows the structural alignment of Subt9195 and 3AFGA. All of the top five enzyme homologs in PDB predict the consensus EC number: 3.4.21.62 and the active site location for Subt9195; Asp-63, His-101, Asn-197 and Ser-262.

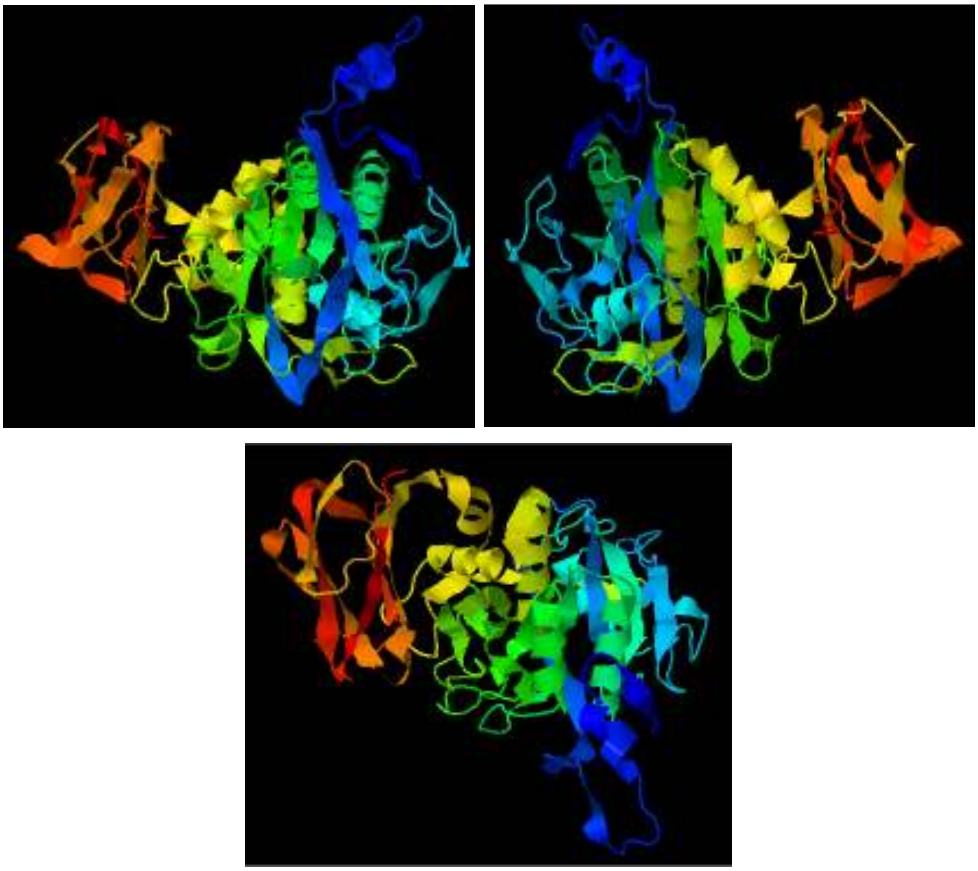


Figure 4.27: Predicted 3D model of Subt9195 from different angle. Front view:(top left), back view (top right) and top view (below) of the. The protein is depicted as a rainbow coloured cartoon; N-terminus=Blue, C terminus=Red

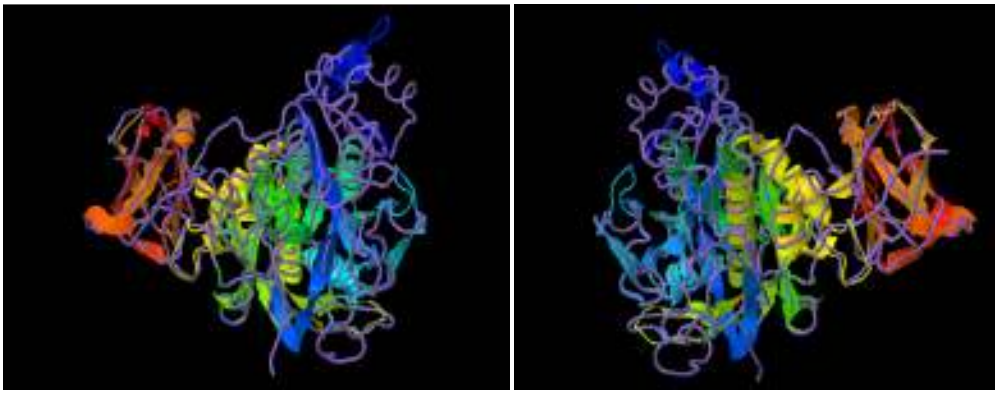


Figure 4.28: Structural alignment of Subt9195 and 3AFGA (left: front view, right: back view). Subt9195 is shown in cartoon, while 3AFGA is displayed using backbone trace.

4.3.3.5 Structure modelling of Subt8715

The predicted model for Subt8715 (Model 1) (Figure 4.29a) from the translated protein sequence of the complete ORF had a low TM score and low number of decoys, which indicated low accuracy and incorrect topology (Table 4.8). When the region that was identified as OmpA-C-terminal-like domain was removed, the new predicted model (Figure 4.29b) showed an increase in the number of decoys suggesting a higher accuracy and correct topology. The highest structural similarity of both models was assigned to fervidolysin of *Fervidobacterium pennivorans* (1R6VA). The second model had a TM score > 0.5, which indicated that it shared a similar fold with its closest structural analog. Figure 4.30 shows structural alignment of Model 1 and 2 with 1R6VA respectively.

Table 4. 8: Estimated accuracy of the structure model

Model	C-score	Exp.TM-Score	Exp.RMSD	No.of decoys	Cluster density
1	-2.04	0.47+-0.15	14.2+-3.8	84	0.0410
2	-1.66	0.51+-0.15	11.3+-4.5	657	0.0560

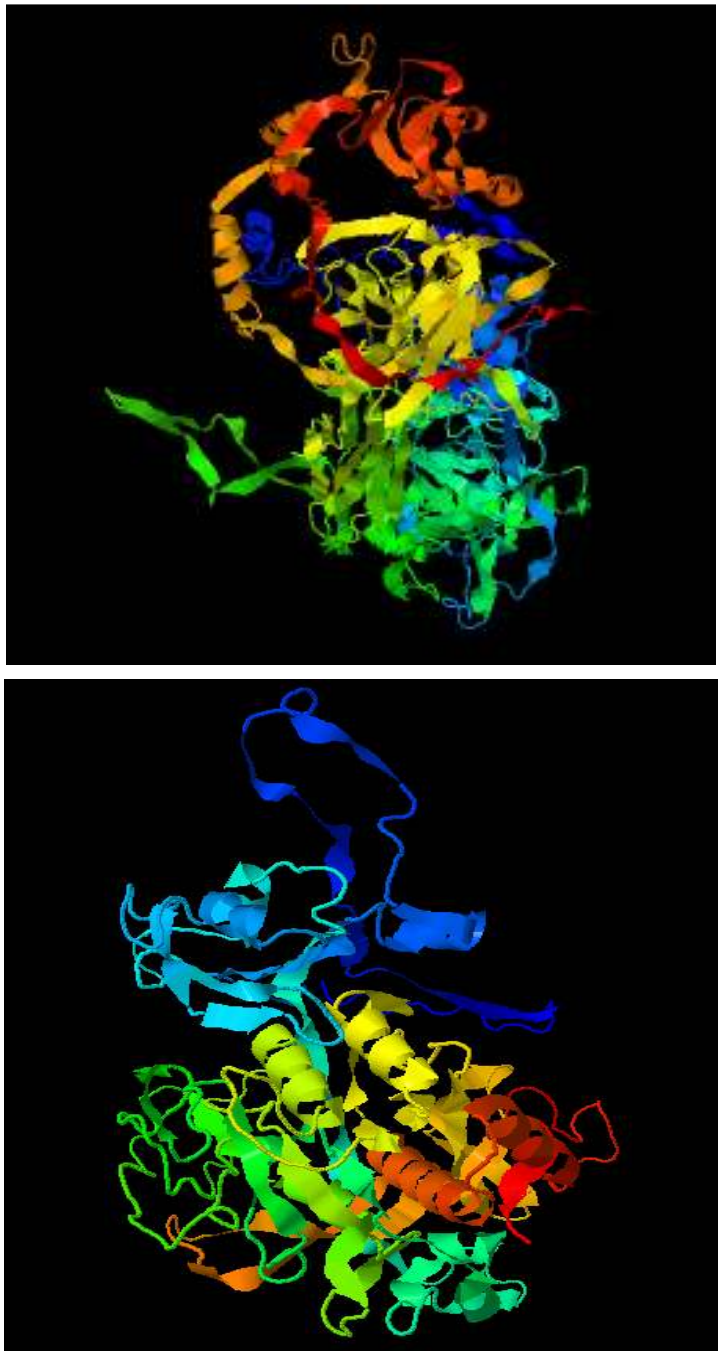


Figure 4.29**:** Predicted 3D model of Subt8715. Model 1 (above); Model 2 (below).

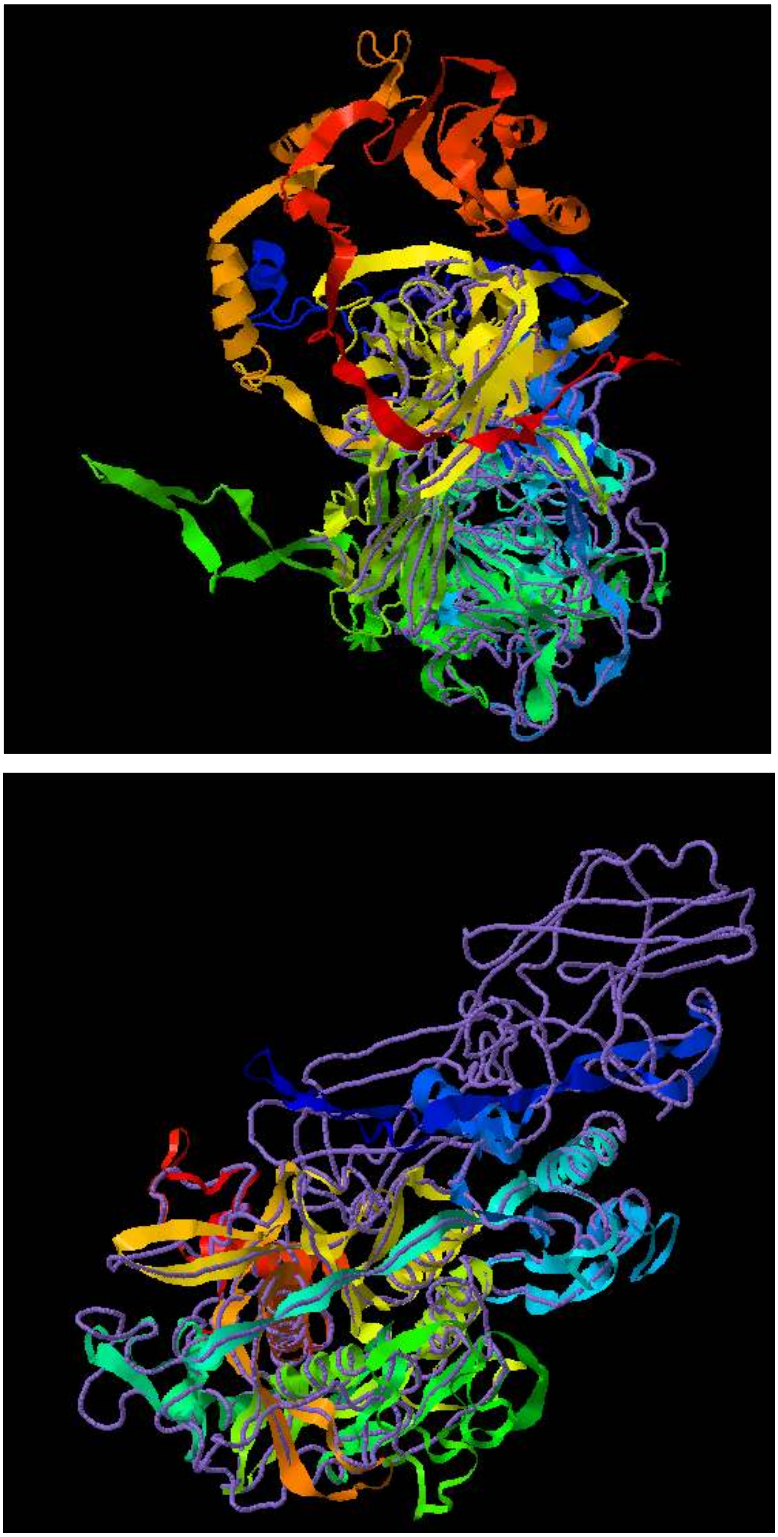


Figure 4.30: Structural alignment of Model 1 (above) and Model 2 (below) with 1R6VA. Query structure is shown in cartoon, while the structural analog is displayed using backbone trace.

4.4 Discussion

4.4.1 Protein properties and overexpression in pET expression system

In this project, four of the subtilase genes from Ace Lake were subjected to cloning and expression studies under pET expression system. Two of the four genes; Subt9195 and Subt5372 were successfully cloned into the pET28b vector but, unfortunately expressed as IB. The attempt to solubilize Subt9195 IB in 8M urea and sarkosyl was successful, however the refolding process to obtain active protease was unsuccessful. In the protease assay using AAPF as substrate, the refolded proteins did not exhibit any protease activity. Since the trials to obtain active enzymes from Subt9195 and Subt5372 were unsuccessful, the study was continued with the cloning and expression study of Subt8715, and Subt4518. In addition, Subt9195 was also cloned as NusA-fusion protein using pET43a, an expression vector that has NusA protein fusion system that is proven with the ability to increase the solubility of recombinant proteins. One of the genes, Subt8715 was successfully cloned and expressed intracellularly as soluble active protease in pET expression system using vector pET28b while the cloning of Subt4518 was unsuccessful.

All the three proteins (Subt9195, Subt5372 and Subt8715) expressed in this project had positive GRAVY values. In two high-throughput expression studies using the *E. coli* expression system (Luan *et al.,* 2004, Madhavan *et al.,* 2010), hydrophobicity was found to be a major factor in determining the success of protein expression (Luan *et al.,* 2004). The GRAVY score (Kyte and Doolittle, 1982), a global descriptor of hydropathy, was used as a reference where more hydrophobic proteins have positive GRAVY values while hydrophilic proteins have negative values (Luan *et al.,* 2004). Interestingly Subt5372 and Subt8715 both have a GRAVY score value of 0.16 but exhibit a different expression pattern. This is in line with the observation which shown that empirical screenings (laboratory experimentation) appears to be the only reliable way to identify soluble expression (Luan *et al.,* 2004). Low GRAVY scores imply expressibility, however, soluble expression depends on other factors and cannot be accurately predicted by bioinformatics alone (Luan *et al.,* 2004).

Expression as IB and unsuccessful cloning of the subtilase genes in pET expression system might be an indicator of the unsuitability of the genes towards this system. Even

though it is known that the pET expression system utilized in this project used a strong T7 promoter to enhance overexpression of the recombinant protein, the strong promoter itself could be the cause of the formation of IB. The *E. coli* machinery is sometimes unable to correctly fold translated proteins, thereby promoting protein aggregation and the formation of IB (Francis and Page, 2001). In some cases, different expression system is suitable for different proteins. For example, one attempt to design a suitable expression system for the *vpr* protease gene involved three expression vectors, i.e., pET, pJOE3075.3 and pBAD TOPO expression vector (Arnórsdóttir *et al.,* 2002). The gene was expressed successfully only in pBAD TOPO system, was unable to be cloned in pET system and formed IB when expressed using the pJOE vector in *E. coli* strain JM109 (Arnórsdóttir *et al.,* 2002).

4.4.2 Extracellular localization and autoprocessing of protease propeptide region

While expression of recombinant protein as extracellular protein is favourable (Mergulhao *et al.,* 2005), the attempts to secrete recombinant proteins somehow face several problems such as incomplete translocation across the inner membrane (Baneyx, 1999) due to insufficient capacity of the export machinery, and proteolytic degradation (Huang *et al.,* 2001). The physico-chemical analysis of Subt9195 and Subt8715 predicted that Subt9195 was an extracellular protein while Subt8715 was predicted as intracellular protein. Observation in the expression studies of both genes indicated that Subt9195 was not expressed successfully as extracellular protein while Subt8715, lacking a signal peptide, has successfully expressed as soluble cytoplasmic protein. The inability of Subt9195 to be expressed as extracellular protein in pET expression system could be explained by many factors. One possible factor is extracellular localization and the autoprocessing steps involved in the production of active protease. Hydrolysis of the peptide bond by protease is an irreversible process, therefore this enzyme, whether secreted or not is always expressed in the form of inactive precursor. The inactive precursor is activated through cascades of controlled complex process only when the active protease is really needed for the cell (Neurath and Walsh, 1976).

Most of the subtilase are extracellular enzymes that are synthesized as their inactive precursor form, called pre-pro-subtilisin, in a cell. This precursor form has a presequence (signal peptide) and a prosequence (propeptide) that are attached to the N-terminal region of

the subtilase mature domain (Siezen and Leunissen 1997). Subtilase are secreted outside the cell as a pro-subtilisin with the assistance of a signal peptide. The pro-subtilisin region was later autoproteolytically removed in the subtilase activation process (Shinde and Inouye, 1996, Shinde and Inouye, 2000).

Subt9195 was cloned with its native signal peptide. It was expected that the native signal peptide would be recognized by the *E. coli* expression system and resulting in the secretion of the Subt9195 protein in the pro-subtilisin form outside of the cell. It was suspected that the formation of IB in the expression studies of Subt9195 was a result of the inability of the host cell to accurately carry out extracellular localization for transport of the precursor protease to the outside of the cell. The SDS-PAGE result showed that Subt9195 was expressed only in the insoluble fraction, no expression band was detected in the extracellular protein fraction. The size of the expressed protein (IB) was ~50 kDa, similar to the size of the predicted molecular weight of Subt9195 (with signal peptide) (Figure 4.21 and 4.24). The size of the expressed Subt9195 indicated that the signal peptide was not successfully removed. Since Subt9195 was not transported to the outside of the cell, further steps of autoprocessing were not able to take place and the inactive Subt9195 was finally expressed as IB. The *E. coli* expression system may not recognize the signal peptide of Subt9195 that was predicted in the bioinformatics analysis as a subtilase gene of uncharacterized Gram-positive Actinobacteria.

There was a possibility that the cloning of Subt9195, together with its native signal peptide, was the reason for the expression of Subt9195 as IB. Similar result was reported by Kwon *et al.* (2011) on performing large scale protein expression studies of 38 protease genes attached to their native signal peptide. The report indicated very low expression success whereby only one protein that was expressed together with the native signal peptide was successfully expressed in soluble form.

4.4.3 Solubilization of Subt9195 protein

Since factors leading to the expression of recombinant protein as IB are not only host-specific, but also target protein specific, no universal remedy for obtaining soluble and active

recombinant protein is available (Cabrita *et al.,* 2006, Hartinger *et al.,* 2010). In the context of obtaining soluble Subt9195 protein, several approaches were explored including expression at reduced temperature, solubilization and *in vitro* refolding of IB as well as the use of solubility-enhancing fusion tag.

Low expression temperature is one approach that can prevent the formation of IB (Schein, 1989). According to Feller *et al.* (1998), a decreased translation rate at low temperatures seems to allow sufficient time for the nascent recombinant protein to fold correctly. Kulakova et al. (1999) and Song et al. (2012) have demonstrated the success to obtain soluble and active recombinant protein at 15 and $\leqslant$ 10°C respectively. However, Subt9195 expressed as IB at 16°C and was not expressed at all at 10°C suggesting that expression conditions for each specific protein must be optimised accordingly..

In another approach, IB of Subt9195 was successfully solubilized in 8M Urea and 0.05% sarkosyl. However the *in vitro* refolding of solubilized Subt9195 to its native state was considered unsuccessful as the assay of the protein using AAPF, after the refolding step, did not show any proteolytic activity. Several cases of successful refolding of active protease from their IB have been reported (Kannan *et al.,* 2001, Wang *et al.,* 2003, Pulido *et al.,* 2006). The main difference between Subt9195 and the other reported cases was that the proteins were expressed as intracellular proteins with a pro-region (without native signal peptide) while in contrast, Subt9195 was expressed with its signal peptide and pro-region in order to obtain extracellular protease. Unsuccessful *in vitro* refolding of Subt9195 as active protease could support our previous argument in the section 4.4.2, that the *E. coli* expression system was not able to recognize and cleave the signal peptide of Subt9195.

Irrespective of the mechanisms, it has been reported that some recombinant proteins that form insoluble aggregates can be expressed in soluble form when tagged with a solubility-enhancing protein (Mitchell *et al.,* 1993, Bach *et al.,* 2001). This route is also favourable as the process of *in vitro* refolding of IB is not a high-throughput option (Nallamsetty and Waugh, 2006). Among the number of solubility-enhancing proteins, NusA, MPB and GST are examples of those that have been experimentally validated and available commercially (Waugh, 2005). The trial to solubilize the IB with high molecular weight fusion protein, NusA, was successful. This was indicated by the presence of expression bands

observed in the SDS-PAGE of the soluble fraction from the expression studies of Subt9195NusAHisC at 30°C (Figure 4.13). Since the gel showed three different sizes of expression band (~100, ~70 and ~50 kDa), the soluble fractions were subjected to partial purification in the affinity system to isolate and identify which of the expression band was the active protease. Subt9195NusAHisC was previously cloned with His-tag at both N and C-terminal ends. This was to make sure that the protein could still be purified by affinity chromatography even after propeptide region removal by autoprocessing. The protease assay of partially purified fraction 4 exhibited protease activity when assayed with AAPF. However, the enzyme showed to be self digesting when stored at 4°C (Figure 4.17).

In the attempt to produce more protein to study the self digesting characteristics of Subt9195NusAHisC, 3 L expression cultures were grown. The purification on the affinity column was later performed on three different batches of cell lysate that were previously freeze-dried using the liquid nitrogen. The trial to scale up the culture volume in the expression and purification steps resulted with compromised enzyme activity. Even though the SDS-PAGE gel indicated that the protein was expressed as soluble NusA-fusion, the purified protein in the scale up experiment did not exhibit autoprocessing that removed itself from NusA protein, as seen in the previous experiment (Figure 4.15 and 4.16). The protease assay, using AAPF as substrate also did not detect any proteolytic activity. In order to rule out the effect of the freeze-drying process towards the protein activity, the scale up process was omitted and the experiment was repeated using 1.0 L cultures and fresh cell lysate. The repeated experiment resulted with similar observation, i.e., an expression band in the SDS-PAGE but no protease activity in the purified fraction 4. The cause of inconsistent proteolytic activity was unknown, but it has hampered the effort to further characterize Subt9195NusAHisC enzyme.

4.4.4 Unsuccessful subtilase maturation mechanism effects towards protease activity

It is noteworthy that *in vivo* subtilases are generally produced as inactive precursors and undergo a stepwise maturation process (signal peptide and the propeptide region removal), which enables precise regulation of activity, spatially and temporally (Khan and James, 1998, Gamble *et al.,* 2011). In the expression study of Subt8715, the gene was expressed as

intracellular protein and showed protease activity in the protease assay using AAPF as substrate. Subt8715 shared 40-50% sequence identity with the extracellular protease but lacked a signal peptide. Therefore it was predicted that Subt8715 would be an intracellular protease. According to Vévodová *et al.* (2010), sequence comparison of intracellular and extracellular susbtilisin protease revealed that, although they maintain a sequence identity of 40–50% with their extracellular counterparts, intracellular subtilisin proteases have a number of distinctive features. Most notably, they lack a signal peptide and propeptide region, but have a shorter N-terminal extension which regulates protease activity via proteolytic processing. Intracellular protease precursor activation must be tightly regulated as an active protease activity in the cell could be destructive (Subbian *et al.,* 2005).

The sequence analysis of Subt8715 did not indicate the presence of short conserved LIPY/F motif in the N-terminal region which was a common characteristic of intracellular subtilisin protease (Vévodová *et al.,* 2010, Gamble *et al.,* 2011). The observation of SDS-PAGE gel of expressed protein suggested that Subt8715 was expressed as an unprocessed ~130 kDa protein in the cytoplasmic fraction (Figure 4.20). Other conformations of Subt8715 were undetected, and the proteolytic activity was shown only in the cytoplasmic fraction. The result indicated that Subt8715 was able to exhibit protease activity without being truncated. There was a report of a high molecular weight intracellular subtilase from *Bacillus subtilis* (Sheehan and Switzer, 1990) that formed *in vivo* as an unprocessed, active protease in stationary cells without any inactive precursor formation. It was suggested that the protease activity *in vivo* was control by an *in situ* protease inhibitor. However, this condition was unlikely to apply to Subt8715 as this protein was expressed in *E. coli* and not in its native host.

Khan and James (1998) discussed that the mechanisms of precursor conversion to active enzymes are diverse in nature, ranging from enzymatic or nonenzymatic cofactors that trigger activation, to a simple change in pH that results in conversion by an autocatalytic mechanism. For example, secreted as a processed protease lacking signal peptide and prodomain, LeSBT1 (73 kDa), a subtilase from tomato plant, required an acidic pH to remove an amino-terminal inhibitory peptide resulting in 68 kDa active protease. The 68 kDa LeSBT1 was found to be inactive at pH 7.0 and above (Janzik *et al.,* 2000). In view of this project, Subt8715 gene was isolated from surface water sample taken from Ace Lake at pH 9 (Lauro

et al., 2011). Even though the optimum pH for Subt8715 is not yet established, the increase of protease activity from pH 7.5 to pH 8.8 is consistent with its native environment. The soluble fraction of Subt8715 used for SDS-PAGE was prepared in buffer pH 7.5. Albeit just an assumption based on LeSBT1 protease that was mentioned previously, there is a possibility that the autocatalysis reaction of propeptide removal was occurring at very slow pace. Therefore, the truncated Subt8715 was not at high enough levels to be visible in the gel.

4.4.5 Effect of temperature towards Subt8715 protease activity

Temperature dependence of crude extract of Subt8715 was determined at pH 8.8 and temperatures ranging from 5 to 60°C (Figure 4.23). Its highest level of activity, using AAPF as the substrate was observed at 50°C, which was 10°C higher compared to cold-adapted SapSh from psychrophilic bacterium *Shewanella* Strain Ac10 (Kulakova *et al.,* 1999) and HP70, a metagenome-derived protease (Ribitsch *et al.,* 2012). Subt8715 highest level of activity at 50°C activity was a unexpected as it was isolated from microorganisms living in cold environments of Ace Lake, Antarctica. However, there is evidence that GroEL protein (a chaperone) of psychrophilic *Pseudoalteromonas* TAC125 is not cold-adapted, but rather well suited to functioning during sudden temperature increases (Piette *et al.,* 2012).

The E_a of the reaction catalysed by Subt8715 was determined from an Arrhenius plot of the values shown in Fig. 4.24 and from the slope of the graph, the calculated E_a of Subt8715 was 36.7kJ/mol. Under different experimental conditions, Kulakova *et al*. (1999) reported the E_a of cold-adapted subtilase, SapSh, to be 41.6 kJ/mol compared to the mesophilic enzyme, subtilisin Carlsberg of 57.5 kJ/mol. The E_a of the reactions catalysed by enzymes from cold-adapted microorganisms are usually lower than the E_a of reactions catalysed by the corresponding enzymes from their mesophilic counterparts (Feller and Gerday, 1997, Lonhienne *et al.,* 2000).

Even though the E_a value of Subt8715 was lower than SapSh and Subtilisin Carlsberg, it was important to acknowledge the use of the crude cell extract of Subt8715 in this work. In contrast, Kulakova *et al*. (1999) used the purified SapSh and Subtilisin Carlsberg to determine the optimum temperature and E_a value. Therefore Subt8715 must be purified in order to

obtain data that is more comparable to other cold-adapted subtilases and its mesophilic counterparts.

4.4.6 Prediction of tertiary structure of Subt9195 and Subt8715

Since the purification and protein crystallization of Subt9195 and Subt8715 were beyond the scope of this chapter, the structure modelling of the expressed proteins were reported in this chapter based on the available bioinformatics data. The structural modelling of Subt9195 has shown that 3AFGA, the subtilisin-like serine protease from *Thermococcus kodakaraensis* is the most structurally similar to Subt9195 (Foophow *et al.,* 2010). The 3AFGA protein structure consists of the N-propeptide domain, the subtilisin-like domain and the β-jelly roll domain (Foophow *et al.,* 2010). Similar to 3AFGA and other members of Subtilase Family, three active site residues, Asp-63, His-101, Asn-197 and Ser-262 was identified in the subtilisin-like α/β domain of Subt9195. However, in comparison to 3AFGA, Subt9195 N-propeptide domain was not as compact as 3AFGA. The C-terminal β-sheets of Subt9195 did not form a stable conformation like the β-jelly roll domain of 3AFGA.

Despite the similarities in the subtilisin-like α/β domain, the association to different taxa might contribute to the difference between archaeal 3AFGA and the predicted bacterial Subt9195 structure. The C-terminal β-jelly roll domain structure of 3AGFA is evident in kexin-like proteases (Holyoak *et al.,* 2003, Henrich *et al.,* 2003, Kobayashi *et al.,* 2009), tomato subtilase 3 (Ottmann *et al.,* 2009) but not in bacterial subtilisins except for Kp-43, from *Bacillus* sp. (Nonaka *et al.,* 2004).

On the other hand, Subt8715 has the highest structural similarity to 1R6VA, a keratinase, fervidolysin from *Fervidobacterium pennivoransas* (Kim *et al.,* 2004). In the structural alignment between Subt8715 and 1R6VA, the propeptide region of Subt8715 was not well aligned to the propeptide of the 1R6VA that is characterized by a globular structure (Kim *et al.,* 2004). The catalytic domain of Subt8715 resembles another subtilisin-like α/β domain. The catalytic triad of Subt8715 consists of Ser454, His282, Asp206 corresponding to Ser389, His208 and Asp170 in 1R6VA. One of the prominent features in Subt8715 is the presence of an extended beta sheet that stretches from the propeptide region and covers the

center of the catalytic domain. In 1R6VA, the overall substrate-binding mode in the catalytic cleft is suggested from the propeptide region crossing the whole active site where the C-terminal of the propeptide binds to the active cleft by forming a β-sheet with catalytic domain (Kim *et al.,* 2004). This interaction most likely defines the overall substrate-binding mode in the catalytic cleft of 1R6VA that favours an extended protein chain, in accord with the known β-keratin (Friedrich and Antranikian, 1996). The similarity between both structures could suggest that Subt8715 has similar substrate preference, a hyphothesis that would be interesting to test.

4.5 Conclusion

The discovery of novel cold-adapted enzymes by expressing proteins identified from sequence-based screening of the metagenomic library dataset still holds promise. Even though no activity was detected in the previous functional screening process (Chapter 3), two out of the four genes showed activity when cloned and overexpressed in pET expression system. Subt9195 expressed as soluble intracellular protein when expressed from pET43a vector but the enzyme activity was inconsistent. Another gene, Subt8715, was expressed as intracellular protein when expressed from the pET28b vector. The optimum temperature of Subt8715 at 50°C was unexpected for an enzyme derived from cold-adapted environment. One of the major challenges in this work was the cloning and expression system chosen for use. The development of a range of easily manipulated cloning and expression systems is of utmost importance to develop a high-throughput screening procedure to identify novel, soluble and active proteases.

CHAPTER 5

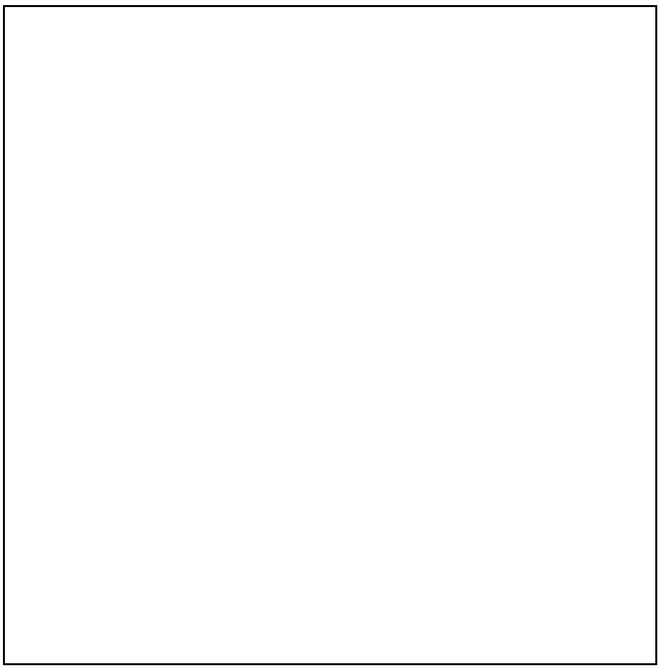
Comparative peptidase analysis of Ace Lake, Organic Lake and Southern Ocean metagenomic data

5.1 Introduction

Microbial utilization of dissolved and particulate organic matter (DOM and POM, respectively) is the nutritional basis of the microbial food web (Azam, 1998). Heterotrophic bacteria mainly catalyze organic matter mineralization (Smith *et al.,* 1992). Polysaccharides and proteins are the most abundant constituents of high molecular weight DOM and are readily utilized by bacteria (Benner *et al.,* 1992). Microbial utilization of polymeric material and POM in aquatic environments depends on the activities of extracellular or cell surface bound hydrolytic enzymes (Smith *et al.,* 1992, Martinez *et al.,* 1996). Peptidases are one of the hydrolytic enzymes that are important in the protein degradation processes. Peptidases can be divided into two large groups according to their substrate specificities, i.e., endopeptidases and exopeptidases (Table 5.1). Exopeptidases remove single or several amino acid residues, dipeptides or tripeptides, from the N- or C-terminal, and accordingly can be classified into mono-, di- and tripeptidases, respectively (Rao *et al.,* 1998). Endopeptidases have been divided into four major groups according to their catalytic mechanism, i.e., serine, cysteine, aspartic and metallo-peptidases (Rawlings and Barrett, 1993, Rao *et al.,* 1998).

Apart of the function of peptidases for nutrition, bacterial membrane-associated peptidase has important functions in the processing, quality control, and regulated turnover of proteins. Quality control and regulated turnover of membrane proteins are necessary not only for the removal of misfolded or damaged proteins in the membrane but also to respond appropriately to stressful environmental conditions (Dalbey *et al.,* 2012). The presence of endopeptidases and exopeptidases in seawater samples have been demonstrated by enzymatic assay of bulk of seawater using various synthetic fluorogenic substrates (Hoppe, 1983, Obayashi and Suzuki, 2005, Obayashi and Suzuki, 2008). While able to illustrate the presence of different proteolytic enzymes in the aquatic environment, this technique is limited by the synthetic substrate applied.

Table 5.1: Classification of peptidase.



Open circles represent the amino acid residues in the polypeptide chain. Solid circles indicate the terminal amino acids, and stars signify the blocked termini. Arrows show the sites of action of the enzyme. Image was taken from (Rao *et al.,* 1998) with modification. Table has been removed due to Copyright restrictions.

More recently, a metagenomics approach has proven useful in providing genetic information on potentially novel biocatalyst or enzymes, linkage between function and phylogeny for uncultured organisms and evolutionary profiles of community function and structure (Thomas *et al*. 2012). The application of this technologies towards sample from the Antarctic lakes and surrounding Southern Ocean has contributed to the wealth of genetic information described in the Ace Lake, Organic Lake and Southern Ocean (Lauro *et al.,* 2011, Yau *et al.,* 2013, Wilkins *et al.,* 2013).

5.1.1 Antarctica: Potential resources of cold-adapted enzymes

While the previous chapters focus particularly on the identification hydrolases and manipulating subtilase genes isolated from Ace Lake, this chapter compares the metagenomic analysis of subtilase and other peptidases from Ace Lake with Organic Lake and the ocean that encircles the continent of Antarctica, i.e., Southern Ocean. Located only a few kilometres from Ace Lake, Organic Lake is a marine-derived hypersaline lake located in the Vestfold Hills, on the eastern shore of Prydz Bay, East Antarctica. It is shallow (~7m) and has variable

surface water temperatures (-14 to +15°C) while remaining sub-zero throughout most of its depth (Franzmann *et al.,* 1987, Gibson, 1999). The salt in the lake was trapped along with the marine biota when the lake was formed (~2700BP) due to the falling of the sea level (~8000BP)(Bird *et al.,* 1991, Zwartz *et al.,* 1998). Through the use of metagenomics, Yau *et al.* (2013) discovered that the Organic Lake ecosystem is dominated by heterotrophic bacteria related to *Psychroflexus* and *Marinobacter*. They also detect the abundance of *Dunaliella*, a unicellular alga (*Chlorophyta)* that has previously been reported to be the dominant alga in the lake (Franzmann *et al.,* 1987).

The stormy seas that encircled the continent of Antarctica are known as the Southern Ocean. As one of the largest marine ecosystems in the world, the Southern Ocean plays a critical role in sustaining marine life around the globe. Upwelling of nutrient rich Circumpolar Deep Water (CDW) returns nutrients transported to the deep ocean by the sinking of organic matter (Rath *et al.,* 1998) and supports 75% of global ocean primary production north of 30°S (Wilkins *et al.,* 2013). Surface waters at high southern latitudes remain cold (3°C) year round but undergo extreme seasonal variations in sea ice cover, light levels and day length (Mitchell *et al.,* 1991, Wilkins *et al.,* 2013). Despite low temperature and seasonal variability that effects productivity, bacteria are abundant in the Southern Ocean and are the major route for carbon flow (Hessen *et al.,* 2004, Brierley and Thomas, 2002). The Southern Ocean is dominated by eukaryotic phytoplankton rather than *Cyanobacteria*, and consists mainly of diatoms, dinoflagellates and haptophytes (Wright *et al.,* 2010, Atkinson *et al.,* 2012). Newcomb Bay is the coastal area of Southern Ocean in East Antarctica. Surface water of Newcomb Bay was highly enriched with *Flavobacteria* in the summer (Williams *et al.,* 2013). Metaproteomics of the surface water sample from Newcomb Bay has revealed the role of *Flavobacteria* in processing algal organic matter in this environment (Williams *et al.,* 2013).

The nature of Ace Lake, Organic Lake and the Southern Ocean, including their location in the permanently cold environment, pH, salinity and the microbial diversity suggested the availability of potential unique cold-adapted enzymes in these environments. In this chapter, a metagenomic dataset of East Antarctica aquatic environments, which include Ace Lake, Organic Lake and Southern Ocean, was analysed for the relative abundance and diversity of subtilase and other peptidases possessed by the uncultivated microbial community in these locations. The associated taxonomic diversities were also analysed in order to infer the physiological roles of the peptidases in the three different cold environments.

5.2 Materials and methods

5.2.1 Data description

The Ace Lake water samples were collected as described in Chapter 2. Organic Lake samples were surface water collected at 68°27' 25.48' S, 78°11' 28.06'E, from the eastern side of the ice free lake (Yau *et al.,* 2011). Southern Ocean samples include surface seawater samples (1-2 m) collected at two locations, 66°16.1'S, 110°32.0'E (235) and 63°52.72'S, 112° 4.2'E (236). Both sites located at 0.4 and 2.5 km from the shore respectively (Williams *et al.,* 2013). Water samples were passed through a 20 µm pore size pre-filter, and microbial biomass was captured by sequential filtration onto 3.0 µm, 0.8 µm and 0.1 µm pore size 293 mm polyethersulfone membrane filters. DNA was extracted from the filters and samples were sequenced using the Roche GS-FLX titanium sequencer (samples from 0.1 µm were also sequenced using Sanger sequencing technology). Reads were processed to remove low quality bases, assembled and annotated as previously described in (Lauro *et al.,* 2011).

Annotated data from 27 assembled datasets that were retrieved from the in-house pipeline, were used as the metagenomic resource in this chapter. Using keyword searches of peptidase, translated ORFs that were annotated as peptidase were analysed. The details of each datasets were summarised in Table 5.2.

Table 5.2**:** List of the datasets used in the metagenomic analysis.

Sample ID	Location	Sampling depth (m)	Filter size (µm)	Total number of reads
236(Open ocean)	Southern Ocean	1-2	0.1	615850
236	Southern Ocean	1-2	0.8	549541
236	Southern Ocean	1.2	3.0	289933
235(Coastal)	Southern Ocean	1-2	0.1	288063
235	Southern Ocean	1-2	0.8	591207
235	Southern Ocean	1-2	3.0	160307
233	Organic lake	n.a	0.1	458368
233	Organic lake	n.a	0.8	489984
233	Organic lake	n.a	3.0	528179
232	Ace Lake	5	0.1	878119
232	Ace Lake	5	0.8	515268
232	Ace Lake	5	3.0	160836
231	Ace Lake	11.5	0.1	863421
231	Ace Lake	11.5	0.8	523321
231	Ace Lake	11.5	3.0	373227
230	Ace Lake	12.7	0.1	540107
230	Ace Lake	12.7	0.8	583495
230	Ace Lake	12.7	3.0	208273
229	Ace Lake	14	0.1	467222
229	Ace Lake	14	0.8	500178
229	Ace Lake	14	3.0	291066
228	Ace Lake	18	0.1	415673
228	Ace Lake	18	0.8	600236
228	Ace Lake	18	3.0	278847
227	Ace Lake	23	0.1	633743
227	Ace Lake	23	0.8	612878
227	Ace Lake	23	3.0	264161

5.2.2 Calculating peptidase associated reads abundance in Ace Lake, Organic Lake and Southern Ocean metagenome dataset

In the assembled data, due to difference of coverage in each scaffold, each annotation might be represented by a different number of reads and therefore a different number of genes in the unassembled data. In order to be able to compare the abundance of peptidases across all samples, the count of reads were adjusted for scaffold coverage as described in (Yutin *et al.,* 2007). The adjustments for each reads were performed by calculating the read equivalents (Equation 1). The read equivalents, r*i,* approximate the number of reads containing each hit in the scaffold.

$$r_i = n_i \cdot g_i / S_i \quad \text{(Equation 1)}$$

Where, n_i is the number of reads in the ith scaffold, g_i is the length of the hits gene fragment on the ith scaffold and S_i is the scaffold length. The sum of read equivalents for each gene will reflects the count of the genes in each dataset.

$$\sum_{i} r_{ij}$$

5.2.3 Data normalization

To account for different levels of sampling across multiple locations, the count for each peptidase were normalized to the mean number of total reads across all samples. The normalization allows comparison of counts between different samples.

5.2.4 Data analysis

Peptidase sequences were analysed according to Clusters of Orthologous Groups (COG)(Tatusov *et al.,* 1997, Tatusov *et al.,* 2003) of proteins annotation. Annotation by KEGG (Kanehisa *et al.,* 2006) and known marker genes (von Mering *et al.,* 2007) were considered for assigning taxonomic and peptidase family classification. Details for each peptidase family were retrieved from MEROPS (Rawlings *et al.,* 2012). Hierarchical clustering and heat-plots were generated with R (R Development Core Team, 2013) using the library 'seriation'. Statistical significance of differences of peptidase associated-COG abundances was assessed using ANOVA test and Tukey-Kramer post-hoc test with confidence intervals at 95% significance, calculated by the Benferroni correction in Statistical Analysis of Metagenomic Profiles (STAMP)(Parks and Beiko, 2010).

5.3 Results

5.3.1 Analysis of COG annotations associated to peptidases in the Ace Lake, Organic Lake and Southern Ocean metagenome

There were total of 163 categories of COG annotations associated to peptidase found in the three locations. Ace Lake was the most diverse with about 144 COG categories while Organic Lake and Southern Ocean have 92 and 93 COG categories each. 72 COG categories that have minimum counts of 100 (relative abundance of 0.20%) in at least one of the sampling location were further analysed in this chapter (Table 5.3) COG0612, COG0739, COG0612, COG0739, COG0465, COG0744, COG0793, COG1770, COG0006, COG1506, COG0308, COG2234, COG0533 and COG1404 were the most abundance COG categories detected.

To examine overall similarity among the datasets, peptidase associated-COG abundance profiles were subjected to hierarchical clustering analysis in a heatmap plot. Using the representation in Figure 5.1, the clustering showed that the COG distributions were neither clustered according to the locations nor the filter size. Southern Ocean samples of 0.1 μm and 0.8 μm, clustered together with Organic Lake sample of 0.8 μm. The plot also indicated that COG distributions in Ace Lake samples clustered according to the upper zones (mixolimnion), interphase, and deep zones (monimolimnion) and further separated into filter sizes. The 0.1 μm Organic Lake sample was shown to be closely related to the 3.0 μm Southern Ocean sample while the 3.0 μm Organic Lake sample was related to Ace Lake mixolimnion 0.1 and 3.0 μm samples.

Similarity test showed that there were significantly different COG categories among the three locations but not among the filter sizes. The most abundance COG categories were not significantly different. As illustrated in Figure 5.2, the categories that were significantly different in abundance ($\rho < 0.05$) were COG0339, COG2936, COG1228, COG3340 and COG2234, which were more abundant in Organic Lake and Southern Ocean samples compared to the Ace Lake.

Table 5.3: COG ID and count of COG annotation related to peptidase from Ace Lake, Organic Lake and Southern Ocean.

COG ID	COGannotations	Peptidase Family	Ace Lake		Organic Lake		Southern Ocean	
			Count	%	Count	%	Count	%
COG0465	ATP-dependent Zn proteases	M41	3739	***8.16***	801	4.74	956	3.67
COG0612	Predicted Zn-dependent peptidases	M16	3500	7.64	772	4.57	1674	***6.43***
COG0739	Membrane proteins related to metalloendopeptidases	M23	3353	7.32	718	4.25	1540	5.91
COG0744	Membrane carboxypeptidase (penicillin-binding protein)	1A	3211	7.01	1083	6.41	1121	4.31
COG0793	Periplasmic protease	S41A	2578	5.63	584	3.46	1208	4.64
COG0006	Xaa-Pro aminopeptidase	M24	1566	3.42	285	1.69	1342	5.15
COG0616	Periplasmic serine proteases (ClpP class)	S49	1396	3.05	103	0.61	259	1
COG0533	Metal-dependent proteases with possible chaperone activity	M22	1257	2.74	126	0.75	455	1.75
COG1770	Protease II	S9A	1240	2.71	1604	***9.5***	815	3.13
COG0260	Leucylaminopeptidase	M17	1196	2.61	51	0.3	215	0.83
COG0024	Methionine aminopeptidase	M24A	1163	2.54	117	0.69	360	1.38
COG0501	Zn-dependent protease with chaperone function	M48	1065	2.32	266	1.57	180	0.69
COG1974	SOS-response transcriptional repressors (RecA-mediated autopeptidases)	S24	966	2.11	100	0.59	239	0.92
COG0681	Signal peptidase I	S26A	899	1.96	177	1.05	445	1.71
COG0308	Aminopeptidase N	M17	879	1.92	552	3.27	1146	4.4
COG0740	Protease subunit of ATP-dependent Clp proteases	S14	847	1.85	29	0.17	86	0.33
COG1686	D-alanyl-D-alanine carboxypeptidase	S11, S12, S13	787	1.72	44	0.26	141	0.54
COG1404	Subtilisin-like serine proteases	S8	775	1.69	365	2.16	697	2.68
COG0768	Cell division protein FtsI/penicillin-binding protein 2	n.a	704	1.54	234	1.39	494	1.9
COG1473	Metal-dependent amidase/aminoacylase/carboxy peptidase	M20D	676	1.47	327	1.94	699	2.68
COG0826	Collagenase and related proteases	U32	658	1.44		0	6	0.02
COG0624	Acetyl ornithine deacetylase /Succinyl-diaminopimelate desuccinylase and related deacylases	M20D	594	1.3	422	2.5	595	2.29
COG1506	Dipeptidyl aminopeptidases /acylaminoacyl-peptidases	S9A	544	1.19	1514	8.97	1013	3.89
COG0542	ATPases with chaperone activity, ATP-binding subunit	S14	526	1.15	283	1.68	30	0.11
COG2027	D-alanyl-D-alanine carboxypeptidase (penicillin-binding protein 4)	S11, S12, S13	519	1.13	31	0.18	5	0.02
COG0405	Gamma-glutamyl transpeptidase	n.a	503	1.1	612	3.62	563	2.16
COG0265	Trypsin-like serine proteases, typically periplasmic, contain C-terminal PDZ domain	S1C	503	1.1	63	0.37	76	0.29
COG0597	Lipoprotein signal peptidase	A8	491	1.07	68	0.4	186	0.72
COG0339	Zn-dependent oligopeptidases	M3	480	1.05	491	2.91	817	3.14
COG0638	Proteasome protease subunit	n.a	472	1.03	18	0.1	120	0.46

Table 5.3: *Continued from previous page.*

COG ID	COGannotations	Peptidase Family	Ace Lake		Organic Lake		Southern Ocean	
			Count	%	Count	%	Count	%
COG1214	Inactive homologs of metal-dependent proteases, putative molecular chaperones	M22	443	0.97	81	0.48	272	1.04
COG2274	ABC-type bacteriocin /lantibiotic exporters, contain an N-terminal double-glycine peptidase domain	n.a	426	0.93	55	0.32	71	0.27
COG1619	Uncharacterized proteins, homologs of microcin C7 resistance protein MccF	U61	420	0.92	110	0.65	268	1.03
COG1026	Predicted Zn-dependent peptidases, insulinase-like	M16C	414	0.9	8	0.05	153	0.59
COG2366	Protein related to penicillin acylase	S45	366	0.8	28	0.16	595	2.29
COG2195	Di- and tripeptidases	M20C	311	0.68	300	1.78	686	2.64
COG0750	Predicted membrane-associated Zn-dependent proteases 1	M50	297	0.65	150	0.89	346	1.33
COG3590	Predicted metalloendopeptidase	M13	283	0.62	457	2.71	403	1.55
COG1680	Beta-lactamase class C and other penicillin-binding proteins	S12	282	0.62	228	1.35	197	0.75
COG0312	Predicted Zn-dependent proteases and their inactivated homologs	U62	278	0.61	37	0.22	35	0.13
COG1989	Signal peptidase, cleaves prepilin-like proteins	A24	271	0.59	9	0.05	52	0.2
COG2738	Predicted Zn-dependent protease	n.a	265	0.58	78	0.46	95	0.36
COG1164	Oligoendopeptidase F	M3	259	0.56	34	0.2	41	0.16
COG1363	Cellulase M and related proteins	M42	255	0.56	164	0.97	311	1.19
COG0466	ATP-dependent Lon protease, bacterial-type	S16	255	0.56	13	0.08	120	0.46
COG2317	Zn-dependent carboxypeptidases	M32	253	0.55	55	0.33	17	0.06
COG1994	Zn-dependent proteases	M50	237	0.52	8	0.05	2	0.01
COG1505	Serine proteases of the peptidase family S9A	S9A	237	0.52	471	2.79	483	1.86
COG2355	Zn-dependent dipeptidase, microsomal dipeptidase homolog	M19	235	0.51	19	0.11	277	1.06
COG2234	Predicted aminopeptidases	M28	227	0.49	781	4.62	1235	4.74
COG2309	Leucyl aminopeptidase (aminopeptidase T)	M29	159	0.35	16	0.09	15	0.06
COG1362	Aspartyl aminopeptidase	M18	149	0.33	7	0.04	88	0.34
COG1876	D-alanyl-D-alanine carboxy peptidase	M15	118	0.26	76	0.45	35	0.13
COG2802	Uncharacterized protein, similar to the N-terminal domain of Lon protease	S16	111	0.24	10	0.06	5	0.02
COG0596	Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily)	S33	109	0.24	32	0.19	279	1.07
COG3291	PKD repeat proteins	n.a	107	0.23		0	6	0.02
COG1132	ABC-type multidrug/protein/lipid transport system, ATPase component	n.a	101	0.22	2	0.01		0
COG3579	Aminopeptidase C	C1	100	0.22		0		0
COG3108	Uncharacterized BCR	n.a	93	0.2	13	0.08		0
COG1220	ATP-dependent protease, ATPase subunit	n.a	85	0.19		0	4	0.02
COG2071	Predicted glutamine amidotransferases	C26	83	0.18	9	0.05	26	0.1

Table 5.3: *Continued from previous page.*

COG ID	COG_annotation	Peptidase Family	Ace Lake		Organic Lake		Southern Ocean	
			Count	%	Count	%	Count	%
COG0791	Cell wall-associated hydrolases (invasion-associated proteins)	Nlp/P60	83	0.18	124	0.73	327	1.26
COG0823	Periplasmic component of the Tol biopolymer transport system	S41	79	0.17	342	2.03	113	0.44
COG3191	L-aminopeptidase/D-esterase	S58	73	0.16	123	0.73	179	0.69
COG1446	Asparaginase	T2	58	0.13	135	0.8	294	1.13
COG3340	Peptidase E	S51	48	0.11	70	0.41	173	0.66
COG2936	Predicted acyl esterases	S15	47	0.1	382	2.26	482	1.85
COG2939	Carboxypeptidase C (cathepsin A)	S10	44	0.1	177	1.05	81	0.31
COG2173	D-alanyl-D-alanine dipeptidase	M15D	36	0.08	6	0.03	171	0.66
COG2267	Lysophospholipase	n.a	6	0.01	63	0.37	27	0.11
COG0441	Threonyl-tRNAsynthetase	S16	4	0.01		0	56	0.22
COG1228	Imidazolonepropionase and related amidohydrolases	n.a	3	0.01	148	0.87	220	0.84

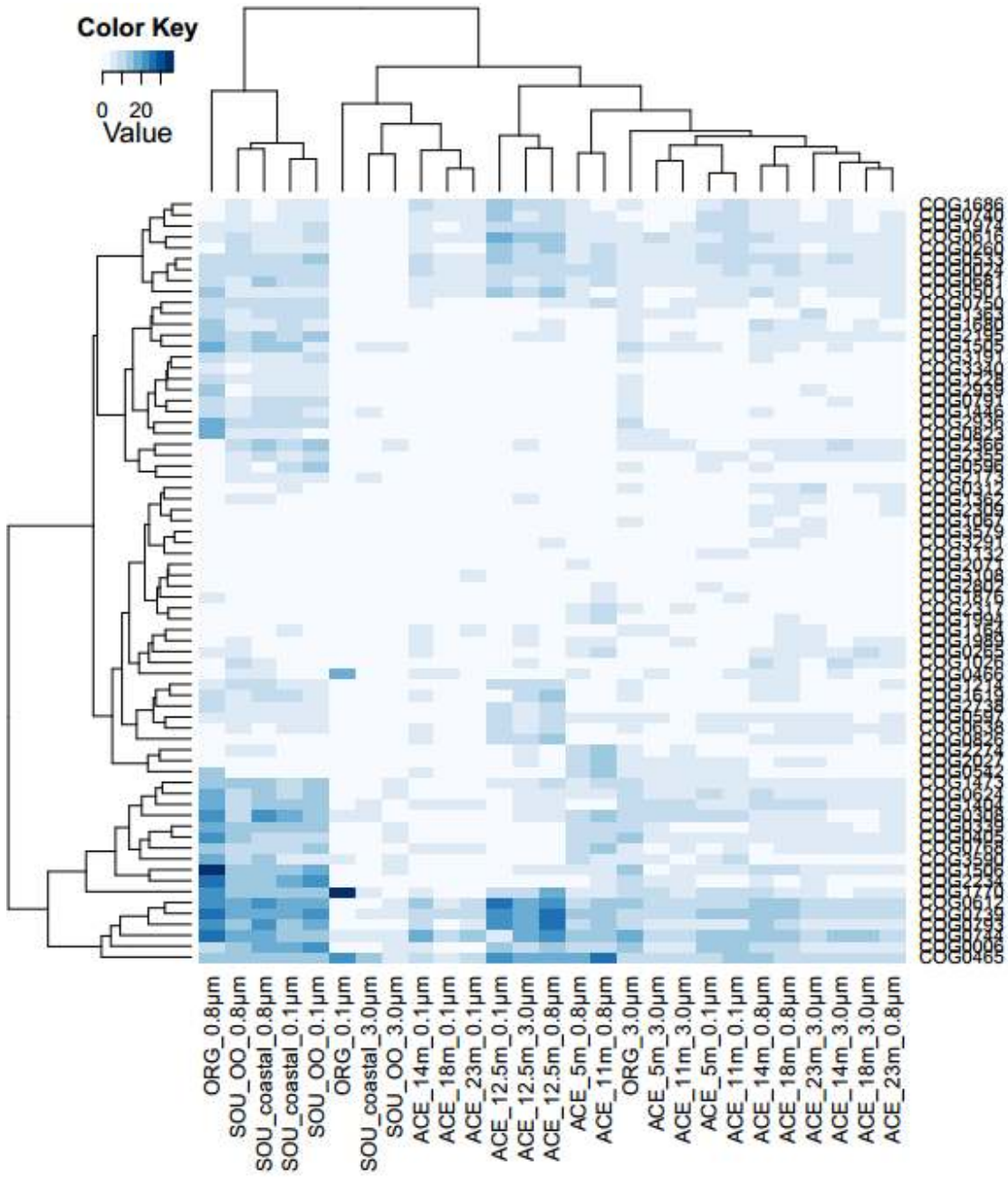


Figure 5.1**:** Hierarchical clustering based on COG category abundances. Values within each category are normalized across samples and square root transformed for better view scale. COG category description were as provided in Table 5.2.

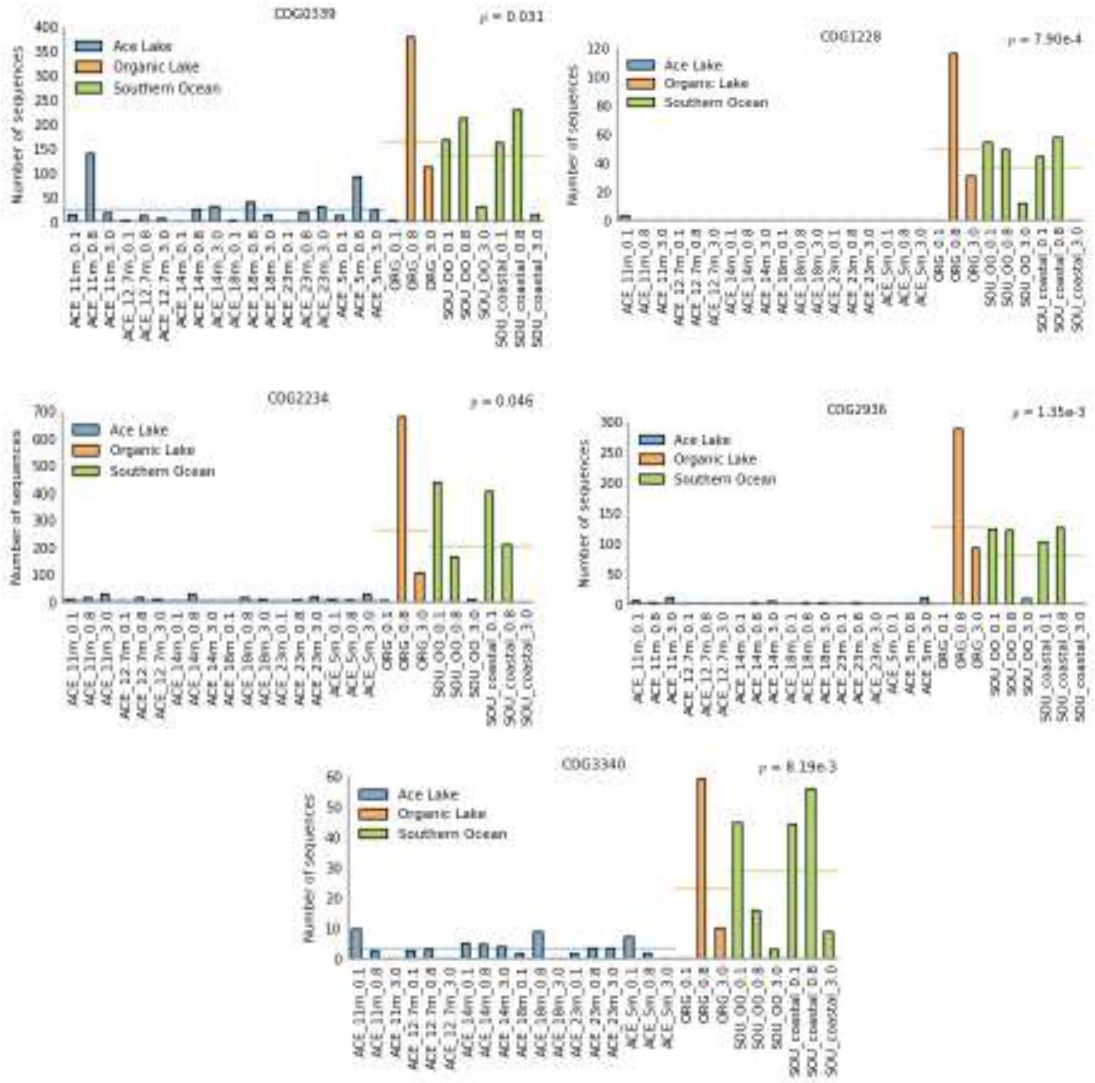


Figure 5.2**:** Statistical (STAMP) analysis of normalized counts of COG annotated proteins among Ace Lake, Organic Lake and Southern Ocean. Only differences with corrected p-value <0.05 are displayed. COG0339: Zn-dependent oligopeptidases, COG1228: Imidazolonepropionase and related amidohydrolases, COG2234: Predicted aminopeptidases, COG2936: Predicted acyl esterases, COG3340: Peptidase E.

5.3.2 Variation of taxa associated with peptidase genes in Ace Lake, Organic Lake and Southern Ocean

To evaluate the diversity of taxonomic groups that associated with peptidase genes, the taxonomic profile of each of the peptidase genes from the three locations were retrieved from KEGG and marker genes annotation, and sorted into phylum, class and genus level (Figure 5.3). The analysis showed that bacteria dominated the population in all three locations at 97.12 to 99.83%. In the Ace Lake, archaea and eucarya made up 1.47% and 1.41% respectively of the population, while in Organic Lake and Southern Ocean samples there were less than 1%. As illustrated in Figure 5.3, *Bacteroidetes* was the dominant phylum in the Southern Ocean and Organic Lake, making up 69.16 and 67.71% each, respectively. On the other hand, in Ace Lake samples the population was dominated by green sulfur bacteria, *Chlorobi* (32.09%), followed by *Proteobacteria* (23.28%) and *Cyanobacteria* (13.39%); *Proteobacteria* made up 24.58% of the peptidase related phyla in Southern Ocean and 16.12% in Organic Lake samples. This analysis also indicated Ace Lake has higher microbial diversity compared to Organic Lake and Southern Ocean.

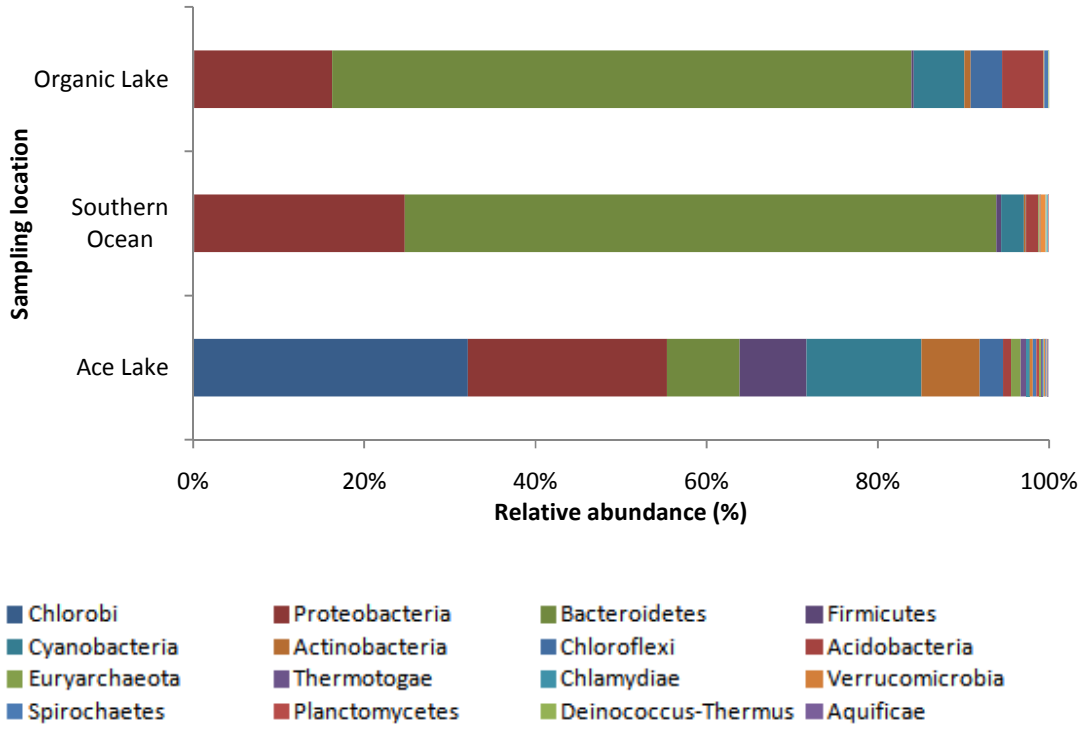


Figure 5.3: Relative abundance of taxa associated with peptidase genes in Ace Lake, Organic Lake and Southern Ocean samples at phyla level.

In order to see the variation of community composition across all sample depths and filter sizes, hierarchical clustering analysis was performed and showed that the samples neither clustered together according to the sampling location or filter size (Figure 5.4). All 0.8 and 0.1 µm Southern Ocean samples clustered together with the 0.8 and 3.0 µm Organic Lake samples. Similar to analysis of COG distributions, Ace Lake sample community composition clustered according to the mixolimnion, interphase, and monimolimnion zones and further systematically clustered according to the filter size except for the 0.8 µm Ace Lake 23 m sample that clustered together with other 3.0 µm samples from the monimolimnion.

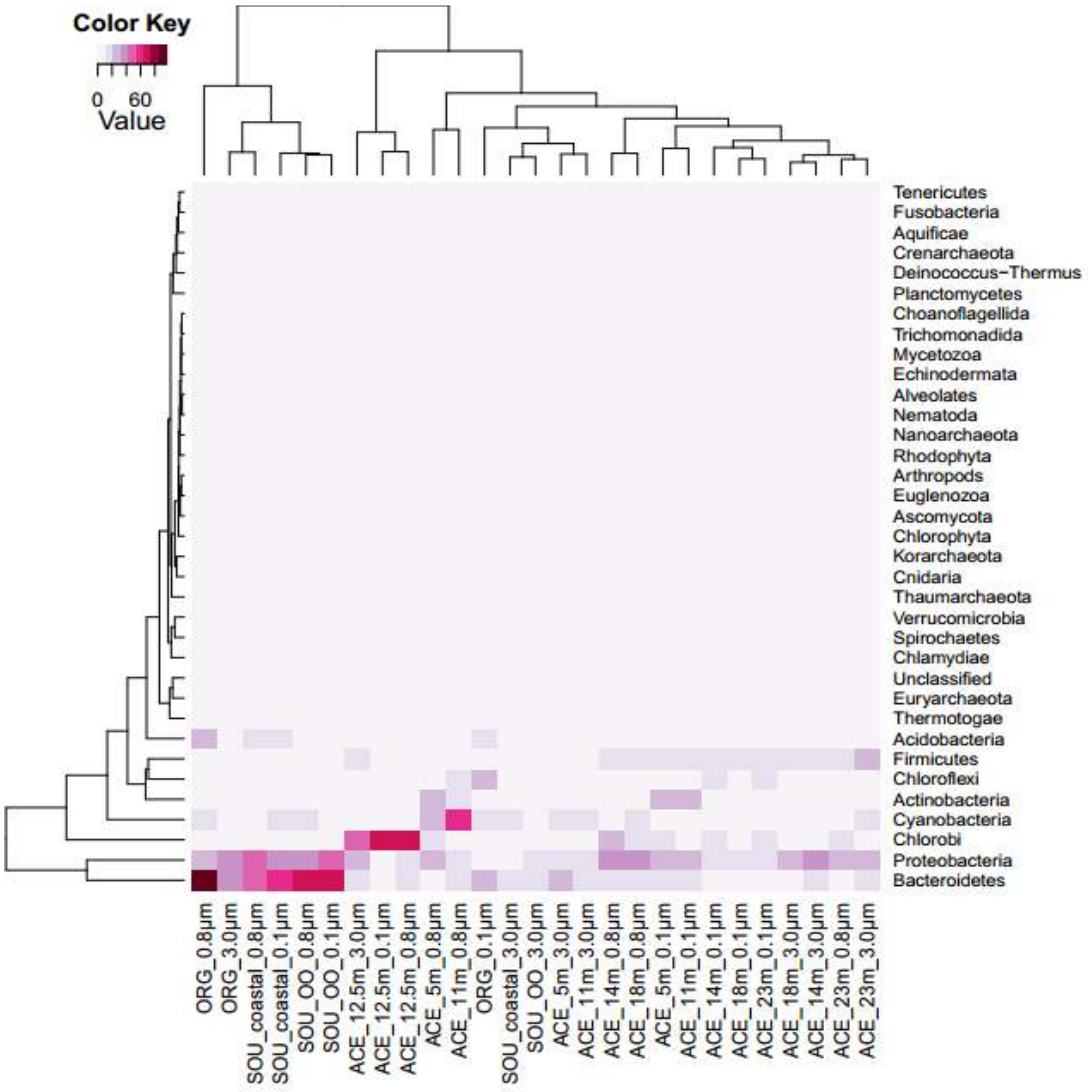


Figure 5.4**:** Hierarchical clustering based on taxonomic associations of KEGGs hits to the peptidase genes. Values within each category are normalized across samples and square root transformed for a better viewscale.ORG: Organic Lake, SOU_coastal: Southern Ocean coastal, SOU_OO: Southern Ocean open ocean, ACE: Ace Lake.

5.3.3 The abundance of metallopeptidase and contribution of the dominant phyla in the sample

As illustrated in Figure 5.5(a), 52% of the COG categories are associated with metallopeptidase, 40% to serine peptidase and 1% to cysteine, aspartic and threonine peptidases. The remaining 5% included peptidases with unknown catalytic classification. The distribution of type of peptidase in the East Antarctica metagenomic dataset was comparable to the Global Ocean Sampling (GOS60) data (Figure 5.5(b)). Both datasets consisted of 40% serine protease. However, in comparison to the East Antarctica dataset, a lower percentage of metallopeptidase (38%) and higher percentage of cysteine peptidase (13%) were detected in the GOS data.

In order to identify the major contributors of metallopeptidase in the East Antarctica samples, the COG category distribution for each phylum detected in each sampling location was determined. As illustrated in Table 5.4, *Chlorobi* were the major contributors of metallopeptidase in the Ace Lake followed by *Cyanobacteria* and *Proteobacteria*. In Organic Lake and Southern Ocean samples *Bacteroidetes* were the major contributors (Table 5.5 and 5.6). For both Organic Lake and Southern Ocean samples, in any COG categories that had low counts or no detectable *Bacteroidetes*, *Proteobacteria* were the major contributors with exception to COG0465 (ATP-dependent Zn proteases) in Organic Lake where *Cyanobacteria* were the major contributors.

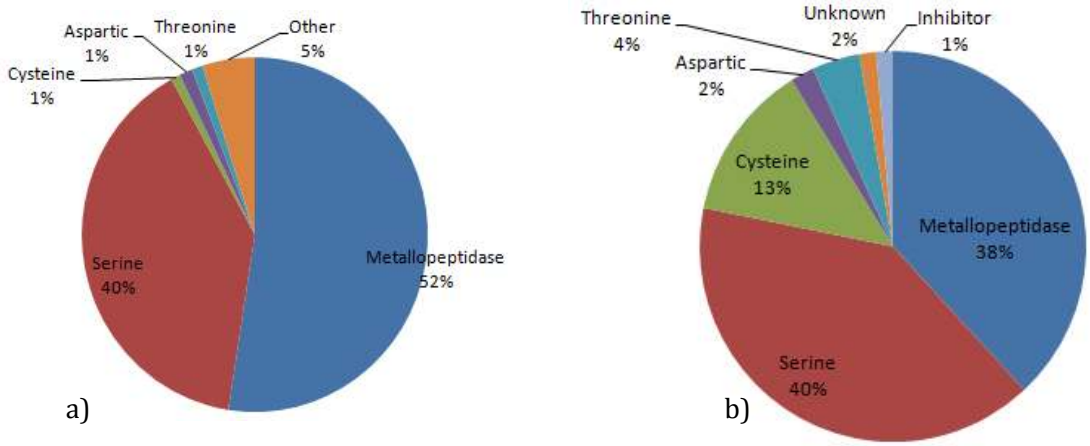


Figure 5.5**:** Comparison of peptidase types distribution in a) Combination of Ace Lake, Organic Lake and Southern Ocean data and b) Global Ocean Sampling expedition samples (GOS60) (Figure b was taken from (Yooseph *et al.,* 2007) with minor modification.

Table 5.4: Contribution of different taxonomic group to counts of COG categories related to metallopeptidase in Ace Lake.

COG ID	COG Description	Acidobacteria	Actinobacteria	Bacteroidetes	Chlorobi	Chloroflexi	Cyanobacteria	Firmicutes	Proteobacteria
COG0612	Predicted Zn-dependent peptidases	24	133	325	***1762***	67	257	251	517
COG0739	Membrane proteins related to metalloendopeptidases	1	135	247	***1572***		253	262	673
COG0465	ATP-dependent Zn proteases	10	210	97	***1416***	79	990	226	614
COG0260	Leucylaminopeptidase		94		***616***	29	135	16	263
COG0501	Zn-dependent protease with chaperone function	7	147	61	***538***	20		38	242
COG0006	Xaa-Pro aminopeptidase	7	180	114	***451***	14	155	124	446
COG0533	Metal-dependent proteases with possible chaperone activity	5	108	88	***385***	10	74	176	287
COG1214	Inactive homologs of metal-dependent proteases, putative molecular chaperones		17	39	***291***	4		49	40
COG1473	Metal-dependent amidase/aminoacylase/carboxypeptidase		36	77	162	2	115	53	***175***
COG2195	Di- and tripeptidases			***117***	39	17		60	67
COG0750	Predicted membrane-associated Zn-dependent proteases 1	2	28	35	2	13	***97***	39	65
COG1876	D-alanyl-D-alanine carboxypeptidase		20				***80***	12	3
COG2234	Predicted aminopeptidases	11		***129***		5			66
COG2173	D-alanyl-D-alanine dipeptidase						26	5	5
COG2355	Zn-dependent dipeptidase, microsomal dipeptidase homolog	8	8	38		2		25	***146***
COG2317	Zn-dependent carboxypeptidases		0	5			***168***	6	23
COG1994	Zn-dependent proteases		2				***139***	47	30
COG0308	Aminopeptidase N		243	133			***346***		138
COG0339	Zn-dependent oligopeptidases			89			***246***		130
COG1362	Aspartylaminopeptidase							98	39
COG2309	Leucyl aminopeptidase (aminopeptidase T)	4	5			13		35	***90***
COG1363	Cellulase M and related proteins		3	93		2	9	19	***96***
COG0624	Acetyl ornithine deacetylase/Succinyl-diaminopimelate desuccinylase and related deacylases	11	162	78		23		74	***176***
COG3590	Predicted metalloendopeptidase	2	***160***	38				3	67
COG1026	Predicted Zn-dependent peptidases, insulinase-like					9	12	6	***369***
COG1164	Oligoendopeptidase F		4	24		14		***65***	60

Counts from the taxonomic group with the greatest contribution to each COG categories are shown in bold and italics.

Table 5.5: Contribution of different taxonomic group to counts of COG categories related to metallopeptidase in Organic Lake.

COG ID	COG description	Acidobacteria	Actinobacteria	Bacteroidetes	Chlorobi	Chloroflexi	Cyanobacteria	Firmicutes	Proteobacteria
COG0612	Predicted Zn-dependent peptidases		6	***687***	3			2	71
COG0739	Membrane proteins related to metalloendopeptidases		1	***593***			6	0	43
COG0465	ATP-dependent Zn proteases		3	256			***492***		37
COG0260	Leucylaminopeptidase		5		5	1			***40***
COG0501	Zn-dependent protease with chaperone function			***208***				5	53
COG0006	Xaa-Pro aminopeptidase		11	***152***		4			109
COG0533	Metal-dependent proteases with possible chaperone activity		4	***100***					22
COG1214	Inactive homologs of metal-dependent proteases, putative molecular chaperones			***74***					7
COG1473	Metal-dependent amidase/aminoacylase/carboxypeptidase		10	***253***			2		57
COG2195	Di- and tripeptidases			***296***					4
COG0750	Predicted membrane-associated Zn-dependent proteases 1		3	***138***					8
COG1876	D-alanyl-D-alanine carboxypeptidase		2				4		***68***
COG2234	Predicted aminopeptidases	193		***435***			14	0	135
COG2173	D-alanyl-D-alanine dipeptidase			2					***3***
COG2355	Zn-dependent dipeptidase, microsomal dipeptidase homolog					2			***16***
COG2317	Zn-dependent carboxypeptidases								***55***
COG1994	Zn-dependent proteases								***8***
COG0308	Aminopeptidase N		5	***493***					53
COG0339	Zn-dependent oligopeptidases			***433***					58
COG1362	Aspartylaminopeptidase								7
COG2309	Leucyl aminopeptidase (aminopeptidase T)					3			
COG1363	Cellulase M and related proteins			***161***					4
COG0624	Acetylornithinedeacetylase/Succinyl-diaminopimelate desuccinylase and related deacylases	4		***263***		2			153
COG3590	Predicted metalloendopeptidase		12	***358***					81
COG1026	Predicted Zn-dependent peptidases, insulinase-like								***8***
COG1164	Oligoendopeptidase F							2	***32***

Counts from the taxonomic group with the greatest contribution to each COG categories is shown in bold and italics

Table 5.6: Contribution of different taxonomic group to counts of COG categories related to metallopeptidase in Southern Ocean.

COG ID	COG Description	Acidobacteria	Actinobacteria	Bacteroidetes	Chlorobi	Chloroflexi	Cyanobacteria	Firmicutes	Proteobacteria
COG0612	Predicted Zn-dependent peptidases			***1548***	7				113
COG0739	Membrane proteins related to metalloendopeptidases			***1182***					348
COG0465	ATP-dependent Zn proteases			***509***			190		256
COG0260	Leucylaminopeptidase								***210***
COG0501	Zn-dependent protease with chaperone function			29	2				***146***
COG0006	Xaa-Pro aminopeptidase	11		617		6		2	***681***
COG0533	Metal-dependent proteases with possible chaperone activity			***242***				2	192
COG1214	Inactive homologs of metal-dependent proteases, putative molecular chaperones			***218***					54
COG1473	Metal-dependent amidase/aminoacylase/carboxypeptidase		14	286				5	***312***
COG2195	Di- and tripeptidases			***686***					1
COG0750	Predicted membrane-associated Zn-dependent proteases 1			***303***	2				41
COG1876	D-alanyl-D-alanine carboxypeptidase								***35***
COG2234	Predicted aminopeptidases	20		***1031***					180
COG2173	D-alanyl-D-alanine dipeptidase	10		23					***138***
COG2355	Zn-dependent dipeptidase, microsomal dipeptidase homolog	46		84					***146***
COG2317	Zn-dependent carboxypeptidases								13
COG1994	Zn-dependent proteases								2
COG0308	Aminopeptidase N	15		***917***					200
COG0339	Zn-dependent oligopeptidases			***601***					212
COG1362	Aspartylaminopeptidase							4	***83***
COG2309	Leucyl aminopeptidase (aminopeptidase T)							5	***8***
COG1363	Cellulase M and related proteins			***307***					3
COG0624	Acetyl ornithine deacetylase/Succinyl-diaminopimelate desuccinylase and related deacylases			***425***					171
COG3590	Predicted metalloendopeptidase	6	1	***334***					63
COG1026	Predicted Zn-dependent peptidases, insulinase-like								***153***
COG1164	Oligoendopeptidase F			3				2	***14***

Counts from the taxonomic group with the greatest contribution to each COG categories are shown in bold and italics

5.3.4 Comparison of COG1404 abundance in the samples

Analysis has shown that COG1404 (Subtilisin), which was the main focus in the previous chapters, was not significantly different among the three locations. In general, subtilisin was most abundant in Ace Lake (775), especially on the 3.0 μm samples followed by Southern Ocean (697) and Organic Lake (365) respectively. In Ace Lake, subtilisin was more abundant in the mixolimnion and monimolimnion compared to the interphase. As indicated in the Table 5.7, *Firmicutes* were the major contributors of subtilisin in Ace Lake followed by *Deltaproteobacteria* and *Bacteroidetes*. In Organic Lake and Southern Ocean, *Bacteroidetes* were the major contributors.

Table 5.7 : Contribution of different taxonomic group to counts of COG1404 (Subtilisin).

Taxon	Ace Lake	Organic Lake	Southern Ocean
Acidobacteria	4	-	-
Actinobacteria	75	4	-
Bacteroidetes	82	180	450
Chloroflexi	62	2	1
Cnidaria	-	-	3
Cyanobacteria	81	143	168
Deinococcus-Thermus	5	-	-
Euryarchaeota	97	-	22
Firmicutes	150	16	16
Amoebozoa	6	-	-
Alphaproteobacteria	2	11	-
Betaproteobacteria	2	-	-
Deltaproteobacteria	136	5	12
Gammaproteobacteria	72	1	13
Thaumarchaeota	-	-	6
Unclassified	1	-	4
Verrucomicrobia	-	3	-

5.4 Discussion

5.4.1 Dominant phyla contribution to the abundance of metallopeptidase

The proportion of peptidase family retrieved in this study was compared to the GOS expedition samples (Yooseph *et al.,* 2007). It was shown that metallopeptidase and serine protease were the most abundant while other peptidases (cysteine, aspartic and threonine peptidases) were detected at lower level in both projects. However, our result was distinguished by the higher abundance of metalloprotease instead of serine protease which dominated the GOS samples. This difference might be explained by the high concentration of zinc in one of the sample subset environment; Ace Lake mixolimnion, which contained zinc at ~79 fold higher compared to the seawater (Rankin *et al.,* 1999, Lauro *et al.,* 2011). The total of normalized counts of COG related to metallopeptidase in the Ace Lake was 66.3 and 36.7% more abundant compared to Organic Lake and Southern Ocean respectively.

The high representation and relative abundant of COG0612, COG0739, COG0465, COG0260, COG0501, COG0793, COG0006, COG0533 and COG1214 in Ace Lake were associated with green sulfur bacterium of the genus *Chlorobium*. These data were consistent with the abundance of *Chlorobium* in the Ace Lake microbial community particularly in the interphase where its cell density was very high (2.2×10^8 cells ml^{-1}) (Ng *et al.,* 2010, Lauro *et al.,* 2011). In Organic Lake, abundance of COGs associated to metallopeptidase were contributed by *Bacteroidetes* of the genus *Flavobacterium*. However, through a metaproteogenomic study of the Organic Lake, analysis of rRNA genes detected over representation of *Bacteroidetes* of a different genus (*Psychroflexus*) in the surface of the lake (Yau *et al.,* 2013). The difference in term of genus classification might be due to the lack of an available reference protein for *Psychroflexu*s in the KEGG database, therefore proteins were assigned to a the much more studied genus *Flavobacterium* (Braun *et al.,* 2005, Duchaud *et al.,* 2007). Southern Ocean metallopeptidase abundance was contributed by the dominant phylum in the samples, *Bacteroidetes*. COG2234 (Predicted aminopeptidase), was a leucyl aminopeptidase linked mainly to *Bacteroidetes* in the Southern Ocean. Leucyl aminopeptidase is a broad substrate specificity exopeptidase with preference for N-terminus Leu, Met or Phe (Jankiewicz and Bielawski, 2003). The activity of leucyl aminopeptidase has been used as an indicator of the potential microbial peptidase activity existing in the ecosystem (Caruso and Zaccone, 2000).

In comparison to another leucyl aminopeptidase (COG0260) that was linked to *Proteobacteria*, the abundance of COG2234 was about five times higher. The abundance of members of *Bacteroidetes* has been reported in the Southern Ocean and associated with phytoplankton blooms (Abell and Bowman, 2005, Williams *et al.,* 2012). Pinhassi *et al.* (2004) reported an increased of aminopeptidase activity associated with the utilization of proteins in a seawater microcosm experiment which had a higher proportion of *Flavobacteria*. More recently, a metaproteomic study of Southern Ocean coastal samples (Newcomb Bay) during summer, detected secreted and cytoplasmic aminopeptidases associated with *Flavobacteria* (Williams *et al.,* 2013) indicating active roles of *Flavobacteria* in decomposition of protein in the aquatic environment.

5.4.2 The prevalence of chaperone, protein regulatory control and membrane biogenesis peptidase in the metagenomic samples

COG0465 (ATP-dependent zinc protease) was the most abundant COG category in Ace Lake. Closer investigation showed that sequences in this category were related to FtsH-like protease. COG0542 and COG0501 which includes Clp and HtpX protein with endopeptidase activity were also abundant in the samples. All these peptidases are important in proteolysis of membrane proteins (Langklotz *et al.,* 2011). Energy-dependent proteases such as Ftsh and Clp perform a crucial role in protein quality control by removing short-lived regulatory proteins and misfolded or damaged polypeptides (Sakoh *et al.,* 2005, Baker and Sauer, 2006, Wagner *et al.,* 2012). In Ace Lake, the abundance of this COG category was due to contributions by *Chlorobi*, *Cyanobacteria*, and *Proteobacteria*.

In Ace Lake, COG0744 (Membrane carboxypeptidase (penicillin-binding protein)) was most abundant in the 0.1 µm monimolimnion samples. In fact, if compared in terms of individual samples, it formed as the most abundant COG in all mixolimnion and monimolimnion 0.1 and 3.0 µm samples. In Organic Lake it was most abundant in the 3.0 µm samples. This protein was detected at low levels in Southern Ocean samples with the highest count in the 3.0 µm open ocean sample. Penicillin-binding proteins (PBP) was responsible for complete assembly of peptidoglycan synthesis performed by a glycosyltransferase that polymerizes the glycan strands, and a transpeptidase activity that cross links the strands via their peptide side chains (Popham and Young, 2003, Sauvage *et al.,* 2008). In this study, COG0744 was detected in bacteria mainly *Chlorobium* and *Proteobacteria* and was not detected in the archaea. The abundance of COG0744 in bacterial lineages is attributed to their ancient evolution as important constituents of the cell wall biosynthesis in bacteria, while it was not detected in archaea probably due to the

different pathways for cell biosynthesis in archaea, which involve pseudomureins (Visweswaran *et al.,* 2011).

Another abundant COG category related to membrane biogenesis was COG0739. Further observation showed that sequences in this category belonged to peptidase family M23, which have endopeptidase activity against a bond within the cross-linking peptide in bacterial cell wall peptidoglycan. COG2027, COG1680, COG0750, COG1876, COG0768 and COG1686 were other membrane biogenesis related proteins detected among the COG categories with relative abundance of more than 0.02% in their respective locations. The abundance of chaperones and peptidases related to protein and membrane regulatory control has been linked to the microbial survival in the ocean surface water, where proteins were continually being damaged as a response to environmental stress exposure (Sowell *et al.,* 2008, Sowell *et al.,* 2011, Williams *et al.,* 2012). Peptidoglycan peptidase also has interesting biotechnological potential in pharmaceutical and cosmetics industry (Maliničová *et al.,* 2010).

5.4.3 Comparison of subtilisin abundance in the samples

Taxonomic diversity of subtilisin in Ace Lake was more diverse compared to Organic Lake and Southern Ocean samples. Major contributors of subtilisin in the Ace Lake were *Proteobacteria* and *Firmicutes*. The analysis in this chapter showed that most of the subtilase associated to *Firmicutes* were detected in the monimolimnion. Even though the counts were less than 100, *Euryarchaeota*, *Bacteroidetes*, *Actinobacteria* and *Chloroflexi* were also detected. With exception of *Firmicutes*, the taxonomic diversity was consistent with the results discussed previously in the Chapter 2, which had used specific profile HMM to search for subtilase in the Ace Lake dataset.

The lowest number of subtilisin detected in the Southern Ocean sample was in the 3.0 µm sample (27) while an almost similar count was detected in the 0.1 (203) and 0.8 µm samples (232). Both 0.1 and 0.8 µm filter were dominated by *Flavobacterium* (*Bacteroidetes*) and *Microcystis* (*Cyanobacteria*). In Organic Lake, subtilisin was very low in the 0.1 µm sample (2), highest in the 0.8 µm sample (290) and about 63 was detected in the 3.0 µm sample. *Cyanobacteria* and *Firmicutes* dominated the 0.8 µm sample while *Bacteroidetes* and *Firmicutes* dominated the 3.0 µm sample. In line with the known subtilisin nonspecific endopeptidase activity, subtilisin in the samples were mostly linked

to phyla that are specialized for the degradation of complex proteins, such as *Bacteroidetes* and *Firmicutes* (Tang *et al.,* 2005, Cottrell and Kirchman, 2000).

5.4.4 Abundance of oligo, di and tripeptidyl peptidase

Abundance of serine oligopeptidase (Family S9) was found in the metagenomic samples. COG1770 (Protease II), which also known as oligopeptidase B were abundant in the Ace Lake mixolimnion and 0.1 µm Southern Ocean coastal sample. Oligopeptidase B is a member of prolyloligopeptidase family that has selectivity towards oligopeptides comprising not more than 30 amino acid residues. Oligopeptidase B cleave Arg and Lys bonds in contrast to other prolyloligopeptidase that cleave prolyl bonds (Polgar, 2002). The analysis showed that *Alphaproteobacteria* of the genus "*Candidatus* Pelagibacter" were the major contributors of oligopeptidase B. This oligotrophic SAR11 clade has a well known characteristic of assimilating free amino acids (Malmstrom *et al.,* 2004), but has also been reported to possess ABC transport systems for oligopeptide uptake (Williams *et al.,* 2012). In the 0.1 µm Organic Lake samples, which had the lowest count of COG categories associated with peptidase, COG 1770 was most abundant and mainly linked to *Gramella* (*Bacteroidetes*), *Chloroflexus* (*Chloroflexi*) and "*Candidatus* Koribacter" (*Acidobacteria*). The result was consistent with a comparative genomic analysis of several *Bacteroidetes* species that showed the prominent role of family S9 prolyloligopeptidases in the peptidase set of *Bacteroidetes* (Bauer *et al.,* 2006). COG0339 and COG1164, oligopeptidase of family M3, most likely involved in intracellular degradation of oligopeptides, were also detected in each of the three sample locations but at lower abundance.

Among the four of di/tripeptidase detected, COG2195 (di- and tripeptidase) was the most abundant compared to COG2936, COG 2173 and COG2355. COG2195 was detected in all three locations and the highest count was detected in Southern Ocean samples, mainly linked to *Bacteroidetes* of the genus *Flavobacterium* This was consistent with transcriptomic studies that link the presence of the transport system and the use of oligopeptides and dipeptides in *Flavobacteria* as source of carbon and nitrogen (Poretsky *et al.,* 2010). COG2936 (Predicted acyl esterases) or the common name X-pro dipeptidyl peptidase, serine peptidase family S15 was detected in low abundance in Ace Lake and in contrast did not link to *Bacteroidetes* of the genus *Gramella*as in Organic Lake and Southern Ocean but associated with *Alphaproteobacteria* and *Acidobacteria*.

5.4.5 Abundance of peptidase E and imidazolonepropionase in the Organic Lake and Southern Ocean

Consistent with the main taxonomic distribution in Organic Lake and Southern Ocean samples, COG 3340 (Peptidase E) was abundant in both locations and linked to *Bacteroidetes* of the genus *Flavobacterium*. Peptidase E was capable of hydrolysing Asp-X peptides and was subsequently shown to have specificity for aspartyl dipeptides (Carter and Miller, 1984, Lassy and Miller, 2000). A study on *Bacteroidetes* derived from the oral cavity, has shown that *Bacteroidetes* favour aspartic acid and its amide asparagine as nutrient source for growth (Shah and Williams, 1987).

Another COG category that was shown to be significantly abundant in Southern Ocean and Organic Lake was COG 1228 (Imidazolone propionate hydrolase(IPase)), a less characterized enzyme in the pathway of histidine degradation to ammonia, glutamate, and one-carbon compounds (formate or formamide) (Bender, 2012). This hydrolase cleaves the urocanic acid ring to yield formiminoglutamate (Bowser Revel and Magasanik, 1958). In Organic Lake and Southern Ocean, sequences encoded for this enzyme was mainly linked to *Bacteroidetes* of genus *Gramella*. The pathway is not well studied in *Bacteroidetes* and the enzyme is predicted based on sequence similarity. However, it is known that this pathway is subject to a strong catabolite repression, preventing bacteria from using L-histidine as a nitrogen source if β-D-glucose or succinate are present in the growth medium (Kimhi and Magasanik, 1970). This enzyme is in very low abundance in Ace Lake, as the dominant taxa in the lake, the green sulfur bacteria (*Chlorobi*) and *Cyanobacteria* appear to lack this pathway (Bender, 2012).

5.5 Conclusion

This chapter describes the first investigation of the comparison of peptidase composition in the Ace Lake, Organic Lake and Eastern Antarctica Southern Ocean via metagenomics approach. Results obtained showed that metallopetidase was the largest peptidase family detected. The most abundance peptidases were chaperone and protein and membrane protein regulatory control peptidase that was linked to microbial survival mechanism towards environmental stress. Additionally, the differences in peptidase composition in each location were identified and showed to be linked to the microbial diversity in the respective locations.

CHAPTER 6

Future perspectives and conclusions

6.1 Introduction

Metagenomics has proven to be a powerful methodology that is currently available to access uncultivable and unique bioresources in extreme environments. This technique, coupled with advances in DNA sequencing technology, has successfully investigated the diversity of Antarctic lakes and Southern Ocean ecosystems. This thesis examined the identification of potential cold-adapted enzymes from Antarctic lakes and Southern Ocean metagenome resources. The most noteworthy contributions of this study were to isolate hydrolase genes from 2,561,323 predicted ORFs of the metagenomic data, develop an expression system for the subtilase gene, and perform a holistic comparison of peptidase gene diversity and abundance in three cold environments (Ace Lake, Organic Lake and the Southern Ocean) with relatively different physico-chemical parameters.

6.2 Possible future work for bioprospecting of enzymes from Antarctic metagenome

The search for hydrolase sequences in the Ace Lake metagenome dataset using specific HMMs (Chapter 2) has identified the complete ORFs of ~60 hydrolases. A similar approach could be used to screen for other type of enzymes. An expression system for several subtilase genes has been successfully developed (Chapter 4) and could be adapted and used for expression studies of other available genes, particularly from the 0.1 μm fractions of the mixolimnion samples that have available DNA clones. The improvement of the strategy to increase the success for bioprospecting novel enzymes from this cold-adapted environment in the future includes the use of high-throughput protein expression and purification platforms as well as the use of multi-host expression system. Other complementary approaches, such as the use of commercially synthesized genes, could be attempted, especially in regard to genes that have no available DNA resources, such as those that were identified via pyrosequencing technologies.

6.2.1 High-throughput expression and purification systems

So far, trial and error using various expression vectors is commonly an integral part of the development of new protein production processes (Balzer *et al.,* 2013). To establish a high-throughput expression and purification system, the recombination-based cloning is an alternative to the simple restriction enzyme-based cloning that was applied in this project (Chapter 4). Recombination-based cloning is mediated by recombinases at site-specific sequences and eliminates the use of restriction enzymes and ligases. The most popular commercial recombination-based cloning systems on the market are the Creator™ Cloning System (Clontech) and the Gateway® Cloning Technology (Life Technologies). Both systems require initial insertion of a gene of interest (GOI) into an entry vector, which may be seen as a drawback of the system because of the addition of many extra steps in the cloning process. However, the entry vector prevents negative selection against toxic genes that may be expressed (due to leaky expression) under a bacterial promoter. Another major advantage is the possibility to transfer GOIs to expression vectors without PCR amplification and consequently without the need to obtain full-length GOI sequences for clone validation (Katzen, 2007).

The availability of Gateway-compatible vectors with variety of solubility fusion tags (Freuler *et al.,* 2008, Caruso and Zaccone, 2000) is an advantage for future work of this project. This is considering the issues of IB formation and the success of NusA solubility-tag usage in producing soluble protein described in Chapter 4. The Gateway® cloning system has the widest adoption and has been employed for several high-throughput ORF cloning projects, including human (Rual *et al.,* 2004, Škalamera *et al.,* 2011), *E. coli* (Rajagopala *et al.,* 2010) and several viruses (Pellet *et al.,* 2010). Factors such as its early release, its ease at handling both N- and C-terminal tags, and the near perfect efficiency of transfer from entry clone to expression vectors, seem to have outweighed the difficulties encountered in creating the entry clone (Festa *et al.,* 2013). Due to the high usage of this system, the collection of Gateway-compatible clones and vectors is the largest among all high-throughput cloning systems and is continually growing (Festa *et al.,* 2013). For example, in the study of recombinant expression and functional analysis of proteases, Kwon and colleagues (2000) shuttled the validated ORFs of interest into multiple Gateway-compatible expression vectors encoding of a variety of fusion tags and achieved high rates of protein solubility and purification.

6.2.2 The advantage of multiple-host metagenomics expression system

The search for hydrolase sequences in the Ace Lake metagenome datasets using specific profile HMMs (Chapter 2), has identified 63 complete ORFs of subtilase, lipase and GH13 family enzymes. Most of the sequences were derived from the 0.1 µm size fraction and were from Sanger-454 hybrid. The sequence-based search also resulted with the detection of genomic signatures for a highly diverse and novel community of anaerobic bacteria and archaea that would be involved in remineralising organic matter in the monimolimnion (Lauro *et al.,* 2011). For example, the one and only complete ORF identified from monimolimnion zone of Ace Lake that was annotated as subtilisin-like serine protease is linked to an anaerobic "*Candidatus* Cloacamonas acidaminovorans", a probable syntrophic bacterium that might derived most of its carbon and energy from the fermentation of amino acids (Pelletier *et al*., 2008). The comparative peptidase analysis of Ace Lake, Organic Lake and Southern Ocean samples described in Chapter 5 also indicated the presence of diverse taxonomic groups in the metagenomic samples. *Bacteroidetes* was the dominant phylum in Southern Ocean and Organic Lake whereas Ace Lake was dominated by green sulfur bacteria (*Chlorobi*).

The taxonomic diversity of the metagenomic sample could be one of the explanations for low positive hits in the agar-based functional screening of the *E. coli*, single-host metagenomic library (Chapter 3). An alternative to the *E. coli* single-host strategy applied in this project is the use of multiple hosts system. This multiple hosts system is able to diversifies the available expression machinery, thus increasing the chance of successful gene expression (Ekkers *et al.,* 2012). Therefore, future studies to obtain novel enzyme from Ace Lake, Organic Lake and Southern Ocean metagenome should consider metagenomic library construction using broad host range shuttle vectors and non-*E. coli* hosts to increase the success of heterologous gene expression. Advanced vector designs commonly incorporate RK2 replicons that allow efficient transfer of the vector from *E. coli* into various other hosts (Aakvik *et al.,* 2009, Kakirde *et al.,* 2011). Comparative screenings of a metagenomic soil library demonstrated that different enzymes can be detected when phylogenetically distinct expression host strains are used (Craig *et al.,* 2010).

Investment in more sophisticated host-vector systems based on a broad range of host organisms is an invaluable alternative to the available single-host system, which has limitation in heterologous expression (Ekkers *et al.* 2012). In particular for Ace Lake metagenomes, the development of host-vector systems for the prevalent strains in this

environment such as *Actinobacteria*, *Bacteroidetes, Acidobacteria, Verrucomicrobia* and *Spirochetes* that are relatively incompatible with the *E. coli* expression system hold great potential to increase the rates of expression of genes from the metagenomes.

6.2.3 The potential success of heterologous expression using synthetic DNA

In the production of recombinant protein, relatively little effort has been made to optimise the genetic components of the expression system compared to the optimisation of fermentation and purification process. Nevertheless, the effort of optimisation is usually directed towards promoters and their associated transcription factors, given the abundance of known and characterized bacterial, archaeal and eukaryotic environment-responsive promoters, which include the well known promoters of the *E. coli lac*, *tet* and *ara* operons (Khalil and Collins, 2010) rather than to the coding sequence of the gene itself (Gustafsson *et al.,* 2012). The situation might be due to the fact that synthetic genes have only been widely and cheaply available for a few years, therefore systematic, well-controlled studies of the relationship between gene design parameters and expression have not been practical before (Gustafsson *et al.* 2004). The current DNA sequencing technologies (that did not require DNA clone library construction) has contributed to the increased of the sequence information from genome and metagenomic projects, but most of the sequence are lacked of available DNA resources (Venter *et al.,* 2004), therefore, promoting the use of genetic construct of synthetic DNA for expression studies (Carlson, 2013). The increase in speed and decrease in cost of synthetic DNA (Figure 6.1) provides a convenient route to obtain genes encoding these virtual proteins (Newcomb *et al.,* 2007). Examples of breakthroughs in synthetic gene manipulation include the group at the JCVI that assembled nearly a megabase of genomic DNA to create *Mycoplasma mycoides* JCVI-syn 1.0 (Gibson *et al.,* 2010), and the success of UC Berkeley's Jay Keasling and colleagues who performed a cutting-edge feat of metabolic engineering by manipulating 12 genes in the artemisinic acid pathway to produced artemisinin in yeast (Ro *et al.,* 2006). Artemisinin is an antimalarial drug commonly extracted from the *Artemisia annua* plant.

It is very interesting that the most abundant peptidases identified in the comparative studies of three cold-adapted environments (Ace Lake, Organic Lake and Southern Ocean) were chaperone and protein and membrane protein regulatory control peptidases that were linked to microbial survival mechanism in response to environmental stresses. For example one of the abundance peptidase group found in the comparative peptidase sequence analysis in Ace Lake, Organic Lake and Southern Ocean

metagenomics data was related to peptidase family M23, which consists of endopeptidases that lyze bacterial cell wall peptidoglycan (Chapter 5). The feature of killing bacterial cells makes the peptidoglycan lyzing enzymes interesting candidates to be use as a prophylactic or therapeutic agent against bacterial infections in humans and animals; as an antimicrobial or enzybiotic for use as a disinfectant in medical, public or private environment; for use as a decontaminant of bacterial contamination in food industry, animal feed or cosmetic industry; or as a general surfactant against bacterially contaminated surfaces (Loessner *et al.,* 2010). The advancement of synthetic DNA studies in a way could complement the study of this group of peptidase that have biotechnological potential in medical, pharmaceutical, and cosmetic industries.

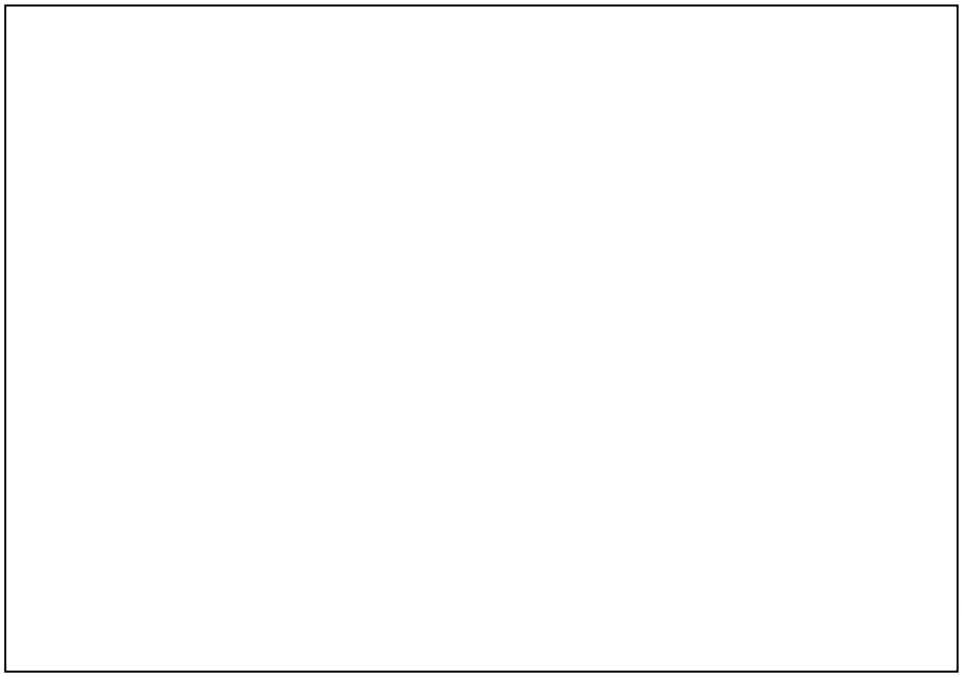


Figure 6.1: Productivity in DNA synthesis and sequencing using commercially available instruments (Carlson, 2013). Figure has been removed due to Copyright restrictions.

The most significant application of synthetic gene technology towards future studies of the kind presented herein is to synthesize candidates genes derived from 454 pyrosequencing that lack DNA clone libraries. Nevertheless, other sequences that have their DNA templates in the clone library are suitable candidates too, as their sequences can be redesigned to improve the success of heterologous expression. Heterologous expression of subtilase, lipases and GH13 genes in *E. coli* might be improved via *de novo* ORF redesign that would include modifications to the existing sequences. This might be most relevant for those genes identified in Chapter 2 that were linked to Gram-positive bacteria and eukaryotes in terms of improving those gene compatibility with available *E.*

coli or non-*E. coli* expression systems. In *de novo* gene design, there can be more than a googol (10^{100}) different ways to encode a specific protein sequence, which is far beyond what can be exhaustively searched by any available technology (Welch *et al.,* 2009b). Companies that provide synthetic gene synthesis have developed their own procedures to accommodate the design and synthesis of optimized gene sequences (as illustrated in Figure 6.2) which include parameters such as codon bias, repeat sequences, mRNA structure and restriction sites (Gustafsson *et al.* 2004).

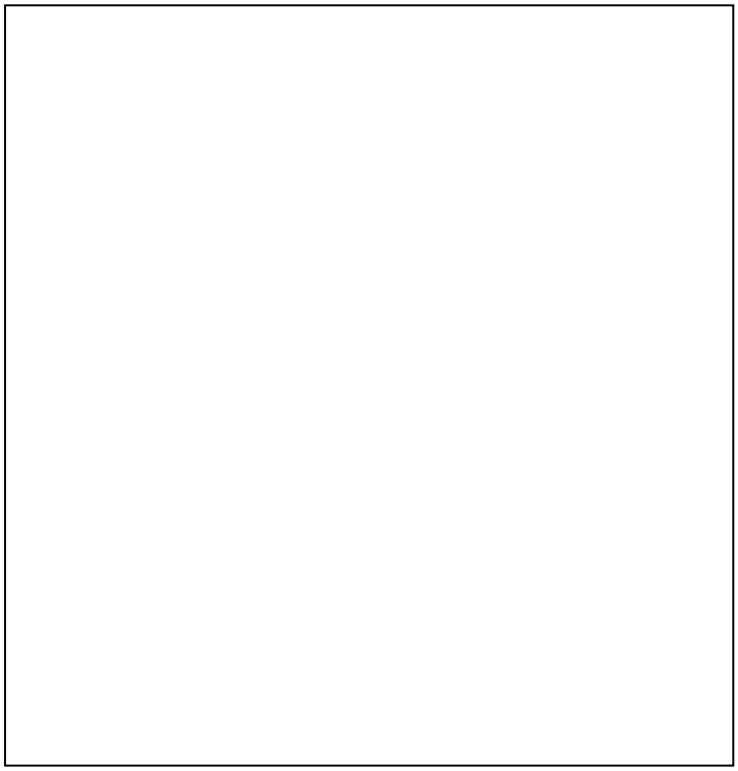


Figure 6.2 : The procedure developed at DNA 2.0 (http://www.dnatwopointo.com/) for designing a gene sequence to encode a specific protein (Gustafsson *et al.,* 2004). Figure has been removed due to Copyright restrictions.

More systematic experimental tests of the effects of sequence-coding variables reported in the past few years are providing data that might force significant revisions of current design assumptions based merely on observations of natural systems. Highly expressed genes in many bacteria and small eukaryotes often have a strong compositional bias, in terms of codon usage. The widely used numerical index, codon adaptation index (CAI) uses this bias to predict the expression level of genes (Sharp and Li, 1987). So far almost all ORF engineering uses design principles based on the codon usage frequency, GC% bias of the expression host organism, and the codon usage of genes that are highly expressed in the expression host (Plotkin *et al.,* 2011). Bias towards codons that are most used in highly expressed native *E. coli* genes (high CAI value) is widely used as the basis of

gene optimisation (Graf *et al.,* 2009). However, as reviewed in Gustafsson *et al.* (2012), recent hypothesis-driven sequence design and testing showed that there is no significant correlation between heterologous protein expression and CAI (Kudla *et al.,*2009, Welch *et al,.* 2009a, Allert *et al.,* 2010).

According to Kudla and colleagues (2009), the majority of the difference in heterologous protein expression can be explained by mRNA folding near the translational start of the ORF. Welch and colleagues (2009a) observed significant correlation between heterologous protein expression and the frequencies of codons used to encode a subset of amino acid. Favourable codons were predominantly those read by tRNAs that are most highly charged (amino-acylated) during amino acid starvation, not codons that are most abundant in highly expressed *E. coli* proteins In another study of heterologous expression of genes designed based on optimum GC % and CAI of *E. coli,* majority of the difference in expression could be attributed to low GC content and low predicted mRNA structure in the 5' end of the ORF (Allert *et al.,* 2010).

Further research is needed to fully understand the experimentally-observed relationships between gene features and expression levels. However, the observed correlations can already serve as the basis for reliable design algorithms as well as providing direction for gene improvement strategies (Gustafsson *et al.,* 2012). Looking ahead, ongoing improvements in the performance of DNA sequencing and synthesis are likely to deliver significant increases in productivity and reductions in cost, thus potentially allowing the extensive mining of genes or metabolic pathways with desirable activity from the Antarctic metagenomic datasets.

6.3 Concluding remarks

The thesis demonstrated the advantage of metagenomics as a molecular technique that provides access to the functional gene repertoire of uncultivated microbial communities in Ace Lake, Organic Lake and the Southern Ocean. The advancements in synthetic gene synthesis and heterologous expression systems discussed in this thesis indicate the possibility of isolation of enzymes with potential biotechnological value. The studies presented here established a foundation for future bioprospecting of enzymes from the Antarctic aquatic environment.

References

AAKVIK, T., DEGNES, K. F., DAHLSRUD, R., SCHMIDT, F., DAM, R., YU, L., VÖLKER, U., ELLINGSEN, T. E. & VALLA, S. 2009. A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbiology Letters,* 296**,** 149-158.

ABDALLAH, A. M., VAN PITTIUS, N. C. G., CHAMPION, P. A. D. G., COX, J., LUIRINK, J., VANDENBROUCKE-GRAULS, C. M. J. E., APPELMELK, B. J. & BITTER, W. 2007. Type VII secretion—mycobacteria show the way. *Nature Reviews Microbiology,* 5**,** 883-891.

ABELL, G. C. & BOWMAN, J. P. 2005. Ecological and biogeographic relationships of class *Flavobacteria* in the Southern Ocean. *FEMS Microbiology Ecology,* 51**,** 265-277.

AKOH, C. C., LEE, G.-C., LIAW, Y.-C., HUANG, T.-H. & SHAW, J.-F. 2004. GDSL family of serine esterases/lipases. *Progress in Lipid Research,* 43**,** 534-552.

ALLEN, M. A., LAURO, F. M., WILLIAMS, T. J., BURG, D., SIDDIQUI, K. S., DE FRANCISCI, D., CHONG, K. W., PILAK, O., CHEW, H. H., DE MAERE, M. Z., TING, L., KATRIB, M., NG, C., SOWERS, K. R., GALPERIN, M. Y., ANDERSON, I. J., IVANOVA, N., DALIN, E., MARTINEZ, M., LAPIDUS, A., HAUSER, L., LAND, M., THOMAS, T. & CAVICCHIOLI, R. 2009. The genome sequence of the psychrophilic archaeon, *Methanococcoides burtonii*: the role of genome evolution in cold adaptation. *The ISME Journal,* 3**,** 1012-1035.

ALLERT, M., COX, J. C. & HELLINGA, H. W. 2010. Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames. *Journal of Molecular Biology,* 402**,** 905-918.

ALVAREZ, M., ZEELEN, J. P., MAINFROID, V., RENTIER-DELRUE, F., MARTIAL, J. A., WYNS, L., WIERENGA, R. K. & MAES, D. 1998. Triose-phosphate Isomerase (TIM) of the Psychrophilic Bacterium *Vibrio marinus* : Kinetic and structural properties. *Journal of Biological Chemistry,* 273**,** 2199-2206.

AMANN, R. I., LUDWIG, W. & SCHLEIFER, K. H. 1995. Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. *Microbiological Reviews,* 59**,** 143-169.

ARNÓRSDÓTTIR, J., KRISTJÁNSSON, M. M. & FICNER, R. 2005. Crystal structure of a subtilisin-like serine proteinase from a psychrotrophic Vibrio species reveals structural aspects of cold adaptation. *FEBS Journal,* 272**,** 832-845.

ARNÓRSDÓTTIR, J., SMÁRADÓTTIR, R. B., MAGNÚSSON, Ó. T., THORBJARNARDÓTTIR, S. H., EGGERTSSON, G. & KRISTJÁNSSON, M. M. 2002. Characterization of a cloned subtilisin-like serine proteinase from a psychrotrophic Vibrio species. *European Journal of Biochemistry,* 269**,** 5536-5546.

ARPIGNY, J. L. & JAEGER, K. E. 1999. Bacterial lipolytic enzymes: classification and properties. *Biochemical Journal,* 343**,** 177-183.

ATKINSON, A., WARD, P., HUNT, B., PAKHOMOV, E. & HOSIE, G. 2012. An overview of Southern Ocean zooplankton data: abundance, biomass, feeding and functional relationships. *CCAMLR Science,* 19**,** 171-218.

AZAM, F. 1998. Microbial Control of Oceanic Carbon Flux: The Plot Thickens. *Science,* 280**,** 694-696.

BACH, H., MAZOR, Y., SHAKY, S., SHOHAM-LEV, A., BERDICHEVSKY, Y., GUTNICK, D. L. & BENHAR, I. 2001. *Escherichia coli* maltose-binding protein as a molecular chaperone for recombinant intracellular cytoplasmic single-chain antibodies. *Journal of Molecular Biology,* 312**,** 79-93.

BAKER, T. A. & SAUER, R. T. 2006. ATP-dependent proteases of bacteria: recognition logic and operating principles. *Trends in Biochemical Sciences,* 31**,** 647-653.

BALZER, S., KUCHAROVA, V., MEGERLE, J., LALE, R., BRAUTASET, T. & VALLA, S. 2013. A comparative analysis of the properties of regulated promoter systems commonly

used for recombinant gene expression in *Escherichia coli*. *Microbial Cell Factories,* 12**,** 26.

BANEYX, F. 1999. Recombinant protein expression in *Escherichia coli*. *Current Opinion in Biotechnology,,* 10**,** 411 - 21.

BANEYX, F. & MUJACIC, M. 2004. Recombinant protein folding and misfolding in *E.coli*. *Nature Biotechnology,* 22**,** 1399-1408.

BAUER, M., KUBE, M., TEELING, H., RICHTER, M., LOMBARDOT, T., ALLERS, E., WÜRDEMANN, C. A., QUAST, C., KUHL, H., KNAUST, F., WOEBKEN, D., BISCHOF, K., MUSSMANN, M., CHOUDHURI, J. V., MEYER, F., REINHARDT, R., AMANN, R. I. & GLÖCKNER, F. O. 2006. Whole genome analysis of the marine Bacteroidetes '*Gramella forsetii*' reveals adaptations to degradation of polymeric organic matter. *Environmental Microbiology,* 8**,** 2201-2213.

BENAROUDJ, N., LEE, D. H. & GOLDBERG, A. L. 2001. Trehalose accumulation during cellular stress protects cells and cellular proteins from damage by oxygen radicals. *Journal of Biological Chemistry,* 276**,** 24261-24267.

BENDER, R. A. 2012. Regulation of the histidine utilization (Hut) system in bacteria. *Microbiology and Molecular Biology Reviews,* 76**,** 565-584.

BENNER, R., PAKULSKI, J. D., MCCARTHY, M., HEDGES, J. I. & HATCHER, P. G. 1992. Bulk Chemical characteristics of dissolved organic matter in the ocean. *Science,* 255**,** 1561-1564.

BERLEMONT, R., DELSAUTE, M., PIPERS, D., D'AMICO, S., FELLER, G., GALLENI, M. & POWER, P. 2009. Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. *The ISME Journal,* 3**,** 1070-1081.

BERLEMONT, R., PIPERS, D., DELSAUTE, M., ANGIONO, F., FELLER, G., GALLENI, M. & POWER, P. 2011. Exploring the Antarctic soil metagenome as a source of novel cold-adapted enzymes and genetic mobile elements. *Revista argentina de microbiología,* 43**,** 94-103.

BIRD, M. I., CHIVAS, A. R., RADNELL, C. J. & BURTON, H. R. 1991. Sedimentological and stable-isotope evolution of lakes in the Vestfold Hills, Antarctica. *Palaeogeography, Palaeoclimatology, Palaeoecology,* 84**,** 109-130.

BLATTNER, F. R., PLUNKETT, G., BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K., MAYHEW, G. F., GREGOR, J., DAVIS, N. W., KIRKPATRICK, H. A., GOEDEN, M. A., ROSE, D. J., MAU, B. & SHAO, Y. 1997. The Complete Genome Sequence of *Escherichia coli* K-12. *Science,* 277**,** 1453-1462.

BORNSCHEUER, U. T. 2002. Microbial carboxyl esterases: classification, properties and application in biocatalysis. *FEMS Microbiology Reviews,* 26**,** 73-81.

BOWMAN, J. P., ABELL, G. & MANCUSO NICHOLS, C. 2005. Psychrophilic Extremophiles from Antarctica: Biodiversity and biotechnological potential. *Ocean and Polar Res,* 27**,** 221-230.

BOWSER REVEL, H. R. & MAGASANIK, B. 1958. The enzymatic degradation of urocanic acid. *Journal of Biological Chemistry,* 233**,** 930-935.

BRADFORD, M. M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry,* 72**,** 248-254.

BRADY, S. F., CHAO, C. J. & CLARDY, J. 2002. New natural product families from an environmental DNA (eDNA) gene cluster. *Journal of the American Chemical Society,* 124**,** 9968-9969.

BRADY, S. F. & CLARDY, J. 2000. Long-chain N-Acyl amino acid antibiotics isolated from heterologously expressed environmental DNA. *Journal of the American Chemical Society,* 122**,** 12903-12904.

BRAUN, T. F., KHUBBAR, M. K., SAFFARINI, D. A. & MCBRIDE, M. J. 2005. *Flavobacterium johnsoniae* gliding motility genes identified by mariner mutagenesis. *Journal of Bacteriology,* 187**,** 6943-6952.

BREEZEE, J., CADY, N. & STALEY, J. T. 2004. Subfreezing growth of the sea ice bacterium "Psychromonas ingrahamii". *Microbial Ecology,* 47**,** 300-304.
BRIERLEY, A. S. & THOMAS, D. N. 2002. Ecology of Southern Ocean pack ice. *Advances in Marine Biology.* Academic Press.
BRONDYK, W. H. 2009. Chapter 11 : Selecting an appropriate method for expressing a recombinant protein. *In:* RICHARD, R. B. & MURRAY, P. D. (eds.) *Methods in Enzymology.* Academic Press.
BROWN, M. V. & BOWMAN, J. P. 2001. A molecular phylogenetic survey of sea-ice microbial communities (SIMCO). *FEMs Microbiology Ecology,* 35**,** 267-275.
BURGESS, R. R. 2009a. Protein precipitation techniques. *Methods in Enzymology,* 463**,** 331-342.
BURGESS, R. R. 2009b. Refolding solubilized inclusion body proteins. *Methods in enzymology,* 463**,** 259-282.
BURKE, C. M. & BUTON, H. R. I988. Photosynthetic bacteria in meromictic lakes and stratified fjords of the Vestfold Hills, Antarctica. *Hydrobiologica,* 165**,** 13-23.
BURTON, H. R. Marine lakes of Antarctica. *In:* TYLER, P., ed. Antarctic Symposium, 1981 University of Tasmania, Hobart, Australia. ANZAAS, 51-58.
CABRITA, L., DAI, W. & BOTTOMLEY, S. 2006. A family of *E. coli* expression vectors for laboratory scale and high throughput soluble protein production. *BMC Biotechnology,* 6**,** 12.
CABRITA, L. D. & BOTTOMLEY, S. P. 2004. Protein expression and refolding – A practical guide to getting the most out of inclusion bodies. *Biotechnology Annual Review.* Elsevier.
CANTAREL, B., COUTINHO, P., RANCUREL, C., BERNARD, T., LOMBARD, V. & HENRISSAT, B. 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research,* 37**,** D233 - D238.
CARLSON, R. 2013. Planning for Toy Story and synthetic biology: It's all about competition.[Online]. Retrieved May 21 from www.synthesis.cc .
CARTER, T. H. & MILLER, C. G. 1984. Aspartate-specific peptidases in *Salmonella typhimurium*: mutants deficient in peptidase E. *Journal of Bacteriology,* 159**,** 453-459.
CARUSO, G. & ZACCONE, R. 2000. Estimates of leucine aminopeptidase activity in different marine and brackish environments. *Journal of Applied Microbiology,* 89**,** 951-959.
CASANUEVA, A., TUFFIN, M., CARY, C. & COWAN, D. A. 2010. Molecular adaptations to psychrophily: the impact of omic technologies. *Trends in microbiology,* 18**,** 374-381.
CAVICCHIOLI, R. 2006. Cold-adapted archaea. *Nature Reviews Microbiology,* 4**,** 331-343.
CAVICCHIOLI, R., CHARLTON, T., ERTAN, H., OMAR, S. M., SIDDIQUI, K. S. & WILLIAMS, T. J. 2011. Biotechnological uses of enzymes from psychrophiles. *Microbial Biotechnology,* 4**,** 449-460.
CAVICCHIOLI, R., SIDDIQUI, K. S., ANDREWS, D. & SOWERS, K. R. 2002. Low-temperature extremophiles and their applications. *Current Opinion in Biotechnology,* 13**,** 253-261.
CHENG, Q., STAFSLIEN, D., PURUSHOTHAMAN, S. S. & CLEARY, P. 2002. The Group B Streptococcal C5a peptidase is both a specific protease and an invasin. *Infection and immunity,* 70**,** 2408-2413.
CHEPYSHKO, H., LAI, C.-P., HUANG, L.-M., LIU, J.-H. & SHAW, J.-F. 2012. Multifunctionality and diversity of GDSL esterase/lipase gene family in rice (Oryza sativa L. japonica) genome: new insights from bioinformatics analysis. *BMC Genomics,* 13**,** 309.
CIEŚLIŃSKI, H., BIAŁKOWSKAA, A., TKACZUK, K., DŁUGOŁECKA, A., KUR, J. & TURKIEWICZ, M. 2009. Identification and molecular modeling of a novel lipase from an Antarctic soil metagenomic library. *Polar Journal of Microbiology,* 58**,** 199-204.

COLLINS, T., HOYOUX, A., DUTRON, A., GEORIS, J., GENOT, B., DAUVRIN, T., ARNAUT, F., GERDAY, C. & FELLER, G. 2006. Use of glycoside hydrolase family 8 xylanases in baking. *Journal of Cereal Science,* 43**,** 79-84.

COTTRELL, M. T. & KIRCHMAN, D. L. 2000. Natural Assemblages of marine proteobacteria and members of the *Cytophaga-Flavobacter* cluster consuming low- and high-molecular-weight dissolved organic matter. *Applied and Environmental Microbiology,* 66**,** 1692-1697.

COWAN, D., MEYER, Q., STAFFORD, W., MUYANGA, S., CAMERON, R. & WITTWER, P. 2005. Metagenomic gene discovery: past, present and future. *Trends in Biotechnology,* 23**,** 321-329.

CRAIG, J. W., CHANG, F.-Y., KIM, J. H., OBIAJULU, S. C. & BRADY, S. F. 2010. Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Applied and Environmental Microbiology,* 76**,** 1633-1641.

CRAIG, J. W., CHERRY, M. A. & BRADY, S. F. 2011. Long-chain N-Acyl amino acid synthases are linked to the putative PEP-CTERM/exosortase protein-sorting system in Gram-negative bacteria. *Journal of Bacteriology,* 193**,** 5707-5715.

D'AMICO, S., COLLINS, T., MARX, J.-C., FELLER, G. & GERDAY, C. 2006. Psychrophilic microorganisms: Challenges for life. *EMBO Rep,* 7.

DALBEY, R. E., WANG, P. & VAN DIJL, J. M. 2012. Membrane Proteases in the bacterial protein secretion and quality control pathway. *Microbiology and Molecular Biology Reviews,* 76**,** 311-330.

DAVIS, G. D., ELISEE, C., NEWHAM, D. M. & HARRISON, R. G. 1999. New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnology and Bioengineering,* 65**,** 382-388.

DE MARCO, V., STIER, G., BLANDIN, S. & DE MARCO, A. 2004. The solubility and stability of recombinant proteins is increased by their fusion to NusA. *Biochem Biophys Res Commun,* 322**,** 766 - 771.

DE PASCALE, D., DE SANTI, C., FU, J. & LANDFALD, B. 2012. The microbial diversity of Polar environments is a fertile ground for bioprospecting. *Marine Genomics,* 8**,** 15-22.

DELONG, E. F. & PACE, N. R. 2001. Environmental diversity of bacteria and archaea. *Systematic Biology,* 50**,** 470-478.

DEMAERE, M. Z., LAURO, F. M., THOMAS, T., YAU, S. & CAVICCHIOLI, R. 2011. Simple high-throughput annotation pipeline (SHAP). *Bioinformatics,* 27**,** 2431-2432.

DEMING, J. W. 2009. Extremophiles: Cold environments. *In:* EDITOR-IN-CHIEF:MOSELIO, S. (ed.) *Encyclopedia of Microbiology (Third Edition).* Oxford: Academic Press.

DUCHAUD, E., BOUSSAHA, M., LOUX, V., BERNARDET, J.-F., MICHEL, C., KEROUAULT, B., MONDOT, S., NICOLAS, P., BOSSY, R. & CARON, C. 2007. Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum*. *Nature Biotechnology,* 25**,** 763-769.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research.,* 32**,** 1792-1797.

EISEN, J. A., COYNE, R. S., WU, M., WU, D., THIAGARAJAN, M., WORTMAN, J. R., BADGER, J. H., REN, Q., AMEDEO, P., JONES, K. M., TALLON, L. J., DELCHER, A. L., SALZBERG, S. L., SILVA, J. C., HAAS, B. J., MAJOROS, W. H., FARZAD, M., CARLTON, J. M., SMITH, R. K., JR., GARG, J., PEARLMAN, R. E., KARRER, K. M., SUN, L., MANNING, G., ELDE, N. C., TURKEWITZ, A. P., ASAI, D. J., WILKES, D. E., WANG, Y., CAI, H., COLLINS, K., STEWART, B. A., LEE, S. R., WILAMOWSKA, K., WEINBERG, Z., RUZZO, W. L., WLOGA, D., GAERTIG, J., FRANKEL, J., TSAO, C.-C., GOROVSKY, M. A., KEELING, P. J., WALLER, R. F., PATRON, N. J., CHERRY, J. M., STOVER, N. A., KRIEGER, C. J., DEL TORO, C., RYDER, H. F., WILLIAMSON, S. C., BARBEAU, R. A., HAMILTON, E. P. & ORIAS, E. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLOS Biology,* 4**,** e286.

EKKERS, D., CRETOIU, M., KIELAK, A. & ELSAS, J. 2012. The great screen anomaly—a new frontier in product discovery through functional metagenomics. *Applied Microbiology and Biotechnology,* 93**,** 1005-1020.
ELBEIN, A. D., PAN, Y. T., PASTUSZAK, I. & CARROLL, D. 2003. New insights on trehalose: a multifunctional molecule. *Glycobiology,* 13**,** 17R-27R.
ELEND, C., SCHMEISSER, C., HOEBENREICH, H., STEELE, H. L. & STREIT, W. R. 2007. Isolation and characterization of a metagenome-derived and cold-active lipase with high stereospecificity for (R)-ibuprofen esters. *Journal of Biotechnology,* 130**,** 370-377.
ERCOLINI, D., RUSSO, F., NASI, A., FERRANTI, P. & VILLANI, F. 2009. Mesophilic and psychrotrophic bacteria from meat and their spoilage potential in vitro and in beef. *Applied and Environmental Microbiology,* 75**,** 1990-2001.
FABER, K. 2011. Biocatalytic Applications. *Biotransformations in Organic Chemistry.* Springer Berlin Heidelberg.
FELLER, G. & GERDAY, C. 1997. Psychrophilic enzymes: molecular basis of cold adaptation. *Cellular and Molecular Life Sciences* 53**,** 830-841.
FELLER, G. & GERDAY, C. 2003. Psychrophilic enzymes: hot topics in cold adaptation. *Nature Reviews Microbiology,* 1**,** 200-208.
FELLER, G., LE BUSSY, O. & GERDAY, C. 1998. Expression of psychrophilic genes in mesophilic hosts: Assessment of the folding state of a recombinant alpha -amylase. *Applied Environmental Microbiology,* 64**,** 1163-1165.
FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution***,** 783-791.
FERRER, M., BELOQUI, A., TIMMIS, K. N. & GOLYSHIN, P. N. 2009. Metagenomics for mining new genetic resources of microbial communities. *Journal of Molecular Microbiology and Biotechnology,* 16**,** 109-123.
FESTA, F., STEEL, J., BIAN, X. & LABAER, J. 2013. High-throughput cloning and expression library creation for functional proteomics. *Proteomics,* 13**,** 1381-1399.
FINN, R., MISTRY, J., TATE, J., COGGILL, P., HEGER, A., POLLINGTON, J., GAVIN, O., GUNASEKARAN, P., CERIC, G., FORSLUND, K., HOLM, L., SONNHAMMER, E., EDDY, S. & BATEMAN, A. 2010. The Pfam protein families database. *Nucleic Acids Research,* 38**,** D211 - D222.
FINN, R., TATE, J., MISTRY, J., COGGILL, P., SAMMUT, S., HOTZ, H., CERIC, G., FORSLUND, K., EDDY, S., SONNHAMMER, E. & BATEMAN 2008. The Pfam protein families database. *Nucleic Acids Research,* 36**,** 281 - 288.
FOOPHOW, T., TANAKA, S., ANGKAWIDJAJA, C., KOGA, Y., TAKANO, K. & KANAYA, S. 2010. Crystal structure of a subtilisin homologue, Tk-SP, from *Thermococcus kodakaraensis*: requirement of a C-terminal beta-jelly roll domain for hyperstability. *Journal of Molecular Biology,* 400**,** 865-77.
FRANCIS, D. M. & PAGE, R. 2001. Strategies to optimize protein expression in *E. coli*. *Current Protocols in Protein Science.* John Wiley & Sons, Inc.
FRANZMANN, P., DEPREZ, P., BURTON, H. & VAN DEN HOFF, J. 1987. Limnology of Organic Lake, Antarctica, a meromictic lake that contains high concentrations of dimethyl sulfide. *Marine and Freshwater Research,* 38**,** 409-417.
FRANZMANN, P. D. & DOBSON, S. J. 1992. Cell wall-less, free-living spirochetes in Antarctica. *FEMS microbiology letters,* 97**,** 289-292.
FRANZMANN, P. D., LIU, Y., BALKWILL, D. L., ALDRICH, H. C., CONWAY DE MACARIO, E. & BOONE, D. R. 1997. *Methanogenium frigidum* sp. nov., a psychrophilic, H_2-using methanogen from Ace Lake, Antarctica. *International Journal of Systematic Bacteriology,* 47**,** 1068-1072.
FRANZMANN, P. D. & ROHDE, M. 1991. An obligately anaerobic, coiled bacterium from Ace Lake, Antarctica. *Journal of General Microbiology,* 137**,** 2191-2196.
FREULER, F., STETTLER, T., MEYERHOFER, M., LEDER, L. & MAYR, L. M. 2008. Development of a novel Gateway-based vector system for efficient, multiparallel

protein expression in *Escherichia coli*. *Protein Expression and Purification,* 59**,** 232-241.

FRIEDRICH, A. B. & ANTRANIKIAN, G. 1996. Keratin degradation by *Fervidobacterium pennavorans*, a novel thermophilic anaerobic species of the order Thermotogales. *Applied and Environmental Microbiology,* 62**,** 2875-2882.

FU, J., LEIROS, H.-K., PASCALE, D., JOHNSON, K., BLENCKE, H.-M. & LANDFALD, B. 2012. Functional and structural studies of a novel cold-adapted esterase from an Arctic intertidal metagenomic library. *Applied Microbiology and Biotechnology***,** 1-14.

FULFORD-SMITH, S. P. & SIKES, E. L. 1996. The evolution of Ace Lake, Antarctica, determined from sedimentary diatom assemblages. *Palaeogeography, Palaeoclimatology, Palaeoecology,* 124**,** 73-86.

GABOR, E. M., ALKEMA, W. B. L. & JANSSEN, D. B. 2004. Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environmental Microbiology,* 6**,** 879-886.

GAMBLE, M., KÜNZE, G., DODSON, E. J., WILSON, K. S. & JONES, D. D. 2011. Regulation of an intracellular subtilisin protease activity by a short propeptide sequence through an original combined dual mechanism. *Proceedings of the National Academy of Sciences,* 108**,** 3536-3541.

GEER, L. Y., DOMRACHEV, M., LIPMAN, D. J. & BRYANT, S. H. 2002. CDART: Protein Homology by domain architecture. *Genome Research,* 12**,** 1619-1623.

GEIGER, O., GONZÁLEZ-SILVA, N., LÓPEZ-LARA, I. M. & SOHLENKAMP, C. 2010. Amino acid-containing membrane lipids in bacteria. *Progress in Lipid Research,* 49**,** 46-60.

GERDAY, C., AITTALEB, M., CHESSA, J. & FELLER, G. 2000. Cold-adapted enzymes: from fundamentals to biotechnology. *Trends in Biotechnology,* 18**,** 103 - 107.

GIBSON, D. G., GLASS, J. I., LARTIGUE, C., NOSKOV, V. N., CHUANG, R.-Y., ALGIRE, M. A., BENDERS, G. A., MONTAGUE, M. G., MA, L., MOODIE, M. M., MERRYMAN, C., VASHEE, S., KRISHNAKUMAR, R., ASSAD-GARCIA, N., ANDREWS-PFANNKOCH, C., DENISOVA, E. A., YOUNG, L., QI, Z.-Q., SEGALL-SHAPIRO, T. H., CALVEY, C. H., PARMAR, P. P., HUTCHISON, C. A., SMITH, H. O. & VENTER, J. C. 2010. Creation of a bacterial cell controlled by a chemically synthesized Genome. *Science,* 329**,** 52-56.

GIBSON, J. A. E. 1999. The meromictic lakes and stratified marine basins of the Vestfold Hills, East Antarctica. *Antarctic Science,* 11**,** 175-192.

GILBERT, J. A., HILL, P. J., DODD, C. E. R. & LAYBOURN-PARRY, J. 2004. Demonstration of antifreeze protein activity in Antarctic lake bacteria. *Microbiology,* 150**,** 171-180.

GILICHINSKY, D., RIVKINA, E., BAKERMANS, C., SHCHERBAKOVA, V., PETROVSKAYA, L., OZERSKAYA, S., IVANUSHKINA, N., KOCHKINA, G., LAURINAVICHUIS, K., PECHERITSINA, S., FATTAKHOVA, R. & TIEDJE, J. M. 2005. Biodiversity of cryopegs in permafrost. *FEMS Microbiology Ecology,* 53**,** 117-128.

GODINHO, L. F., REIS, C. R., TEPPER, P. G., POELARENDS, G. J. & QUAX, W. J. 2011. Discovery of an *Escherichia coli* esterase with high activity and enantioselectivity toward 1,2-O-Isopropylideneglycerol esters. *Applied and Environmental Microbiology,* 77**,** 6094-6099.

GOLDBERG, S. M. D., JOHNSON, J., BUSAM, D., FELDBLYUM, T., FERRIERA, S., FRIEDMAN, R., HALPERN, A., KHOURI, H., KRAVITZ, S. A., LAURO, F. M., LI, K., ROGERS, Y.-H., STRAUSBERG, R., SUTTON, G., TALLON, L., THOMAS, T., VENTER, E., FRAZIER, M. & VENTER, J. C. 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences,* 103**,** 11240-11245.

GOMES, J. & STEINER, W. 2004. The biocatalytic potential of extremophiles and extremozymes *Food Technology and Biotechnology,* 42**,** 223-235.

GONZALEZ-MONTALBAN, N., GARCIA-FRUITOS, E. & VILLAVERDE, A. 2007. Recombinant protein solubility-does more mean better? *Nature Biotechnology,* 25**,** 718-720.

GRAF, M., SCHOEDL, T. & WAGNER, R. 2009. Rationales of gene design and de novo gene construction. *Systems Biology and Synthetic Biology.* John Wiley & Sons, Inc.

GRÄSLUND, S., NORDLUND, P., WEIGELT, J., BRAY, J., GILEADI, O., KNAPP, S., OPPERMANN, U., ARROWSMITH, C., HUI, R. & MING, J. 2008. Protein production and purification. *Nature Methods,* 5**,** 135-146.

GREEN, B. D. & KELLER, M. 2006. Capturing the uncultivated majority. *Current Opinion in Biotechnology,* 17**,** 236-240.

GRIFFIN, D. W. 2008. Non-spore forming eubacteria isolated at an altitude of 20,000 m in Earth's atmosphere: extended incubation periods needed for culture-based assays. *Aerobiologia,* 24**,** 19-25.

GRIZOT, S. & BUCHANAN, S. K. 2004. Structure of the OmpA-like domain of RmpM from *Neisseria meningitidis*. *Molecular Microbiology,* 51**,** 1027-1037.

GUNSTONE, F. D. 1999. Enzymes as biocatalysts in the modification of natural lipids. *Journal of the Science of Food and Agriculture,* 79**,** 1535-1549.

GUSTAFSSON, C., MINSHULL, J., GOVINDARAJAN, S., NESS, J., VILLALOBOS, A. & WELCH, M. 2012. Engineering genes for predictable protein expression. *Protein Expression and Purification,* 83**,** 37-46.

HAHN, M. W., LÜNSDORF, H., WU, Q., SCHAUER, M., HÖFLE, M. G., BOENIGK, J. & STADLER, P. 2003. Isolation of novel ultramicrobacteria classified as actinobacteria from five freshwater habitats in Europe and Asia. *Applied and Environmental Microbiology,* 69**,** 1442-1451.

HANDELSMAN, J. 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews,* 68**,** 669-685.

HANNAH, H. 2008. Polar biotech. *Nature Biotechnology,* 26**,** 1204-1204.

HARDEMAN, F. & SJOLING, S. 2007. Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. *FEMS Microbiology Ecology,* 59**,** 524-534.

HARTINGER, D., HEINL, S., SCHWARTZ, H., GRABHERR, R., SCHATZMAYR, G., HALTRICH, D. & MOLL, W.-D. 2010. Enhancement of solubility in *Escherichia coli* and purification of an aminotransferase from *Sphingopyxis* sp. MTA144 for deamination of hydrolyzed fumonisin B1. *Microbial Cell Factories,* 9**,** 62.

HEATH, C., HU, X. P., CARY, S. C. & COWAN, D. 2009. Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from Antarctic desert soil. *Applied Environmental Microbiology,* 75**,** 4657-4659.

HENIKOFF, S., GREENE, E. A., PIETROKOVSKI, S., BORK, P., ATTWOOD, T. K. & HOOD, L. 1997. Gene families: the taxonomy of protein paralogs and chimeras. *Science,* 278**,** 609-614.

HENNE, A., SCHMITZ, R. A., BOMEKE, M., GOTTSCHALK, G. & DANIEL, R. 2000. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Applied Environmental Microbiology,* 66**,** 3113-3116.

HENRICH, S., CAMERON, A., BOURENKOV, G. P., KIEFERSAUER, R., HUBER, R., LINDBERG, I., BODE, W. & THAN, M. E. 2003. The crystal structure of the proprotein processing proteinase furin explains its stringent specificity. *Nature Structural & Molecular Biology,* 10**,** 520-526.

HESSEN, D. O., ÅGREN, G. I., ANDERSON, T. R., ELSER, J. J. & DE RUITER, P. C. 2004. Carbon sequestration in ecosystems: the role of stoichiometry. *Ecology,* 85**,** 1179-1192.

HOLYOAK, T., WILSON, M. A., FENN, T. D., KETTNER, C. A., PETSKO, G. A., FULLER, R. S. & RINGE, D. 2003. 2.4 Å resolution crystal structure of the prototypical hormone-processing protease Kex2 in complex with an Ala-Lys-Arg boronic acid inhibitor. *Biochemistry,* 42**,** 6709-6718.

HOOVER, R. B. & PIKUTA, E. V. 2010. Psychrophilic and psychrotolerant microbial extremophiles in Polar environments. *Polar Microbiology***,** 115-156.

HOPPE, H. 1983. Significance of exoenzymatic activities in the ecology of brackish water: measurements by means of methylumbelliferyl-substrates. *Marine Ecology Progress Series,* 11**,** 299-308.

HORIKOSHI, K. 1999. Alkaliphiles: Some applications of their products for biotechnology. *Microbiology and Molecular Biology Reviews,* 63**,** 735-750.
HORŇÁK, K., MAŠÍN, M., JEZBERA, J., BETTAREL, Y., NEDOMA, J., SIME-NGANDO, T. & ŠIMEK, K. 2005. Effects of decreased resource availability, protozoan grazing and viral impact on a structure of bacterioplankton assemblage in a canyon-shaped reservoir. *FEMS Microbiology Ecology,* 52**,** 315-327.
HORVATHOVA, V., JANECEK, S. & STURDIK, E. 2001. Amylolytic enzymes: molecular aspects of their properties. *General physiology and biophysics,* 20**,** 7-32.
HOUDE, A., KADEMI, A. & LEBLANC, D. 2004. Lipases and their industrial applications. *Applied Biochemistry and Biotechnology,* 118**,** 155-170.
HOYOUX, A., BLAISE, V., COLLINS, T., D'AMICO, S., GRATIA, E., HUSTON, A. L., MARX, J.-C., SONAN, G., ZENG, Y., FELLER, G. & GERDAY, C. 2004. Extreme catalysts from low-temperature environments. *Journal of Bioscience and Bioengineering,* 98**,** 317-330.
HOYOUX, A., JENNES, I., DUBOIS, P., GENICOT, S., DUBAIL, F., FRANCOIS, J., BAISE, E., FELLER, G. & GERDAY, C. 2001. Cold-adapted beta-galactosidase from the Antarctic psychrophile *Pseudoalteromonas haloplanktis*. *Applied Environmental Microbiology,* 67**,** 1529 - 1535.
HU, J., LI, H., CAO, L., WU, P., ZHANG, C., SANG, S., ZHANG, X., CHEN, M., LU, J. & LIU, Y. 2007. Molecular cloning and characterization of the gene encoding cold-active beta-galactosidase from a psychrotrophic and halotolerant *Planococcus* sp. L4. *Journal of Agriculture and Food Chemistry,* 55**,** 2217 - 2224.
HUGENHOLTZ, P., GOEBEL, B. M. & PACE, N. R. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology,* 180**,** 4765-4774.
INGRAHAM, J. & STOKES, J. 1959. Psychrophilic bacteria. *Bacteriological Reviews,* 23**,** 97-108.
INOUYE, M. 1990. Intramolecular chaperone: the role of the pro-peptide in protein folding. *Enzyme,* 45**,** 314-321.
IQBAL, H. A., FENG, Z. & BRADY, S. F. 2012. Biocatalysts and small molecule products from metagenomic studies. *Current Opinion in Chemical Biology,* 16**,** 109-116.
JACOBS, M., ELIASSON, M., UHLÉN, M. & FLOCK, J. I. 1985. Cloning, sequencing and expression of subtilisin Carlsberg from *Bacillus licheniformis*. *Nucleic Acids Research,* 13**,** 8913-8926.
JAEGER, K.-E. & EGGERT, T. 2002. Lipases for biotechnology. *Current Opinion in Biotechnology,* 13**,** 390-397.
JANEČEK, Š. 2002. How many conserved sequence regions are there in the α-amylase family. *Biologia,* 57**,** 29-41.
JANECEK, S., SVENSSON, B. & HENRISSAT, B. 1997. Domain evolution in the α-amylase family. *Journal of molecular evolution,* 45**,** 322-331.
JANKIEWICZ, U. & BIELAWSKI, W. 2003. The properties and functions of bacterial aminopeptidases. *Acta Microbiologica Polonica,* 52**,** 217-231.
JANZIK, I., MACHEROUX, P., AMRHEIN, N. & SCHALLER, A. 2000. LeSBT1, a Subtilase from tomato plants: Overexpression in insect cells, purification and characterization. *Journal of Biological Chemistry,* 275**,** 5193-5199.
JEON, J., KIM, J.-T., KANG, S., LEE, J.-H. & KIM, S.-J. 2009a. Characterization and its potential application of two esterases derived from the Arctic sediment metagenome. *Marine Biotechnology,* 11**,** 307-316.
JEON, J., KIM, J.-T., KIM, Y., KIM, H.-K., LEE, H., KANG, S., KIM, S.-J. & LEE, J.-H. 2009b. Cloning and characterization of a new cold-active lipase from a deep-sea sediment metagenome. *Applied Microbiology and Biotechnology,* 81**,** 865-874.
JESPERSEN, H., ANN MACGREGOR, E., HENRISSAT, B., SIERKS, M. & SVENSSON, B. 1993. Starch- and glycogen-debranching and branching enzymes: Prediction of structural features of the catalytic (β/α)8-barrel domain and evolutionary relationship to other amylolytic enzymes. *Journal of Protein Chemistry,* 12**,** 791-805.

JIANG, C., LI, S.-X., LUO, F.-F., JIN, K., WANG, Q., HAO, Z.-Y., WU, L.-L., ZHAO, G.-C., MA, G.-F., SHEN, P.-H., TANG, X.-L. & WU, B. 2011. Biochemical characterization of two novel β-glucosidase genes by metagenome expression cloning. *Bioresource Technology,* 102**,** 3272-3278.

JONES, B. V., SUN, F. & MARCHESI, J. R. 2007. Using skimmed milk agar to functionally screen a gut metagenomic library for proteases may lead to false positives. *Letters in Applied Microbiology,* 45**,** 418-420.

KAKIRDE, K. S., WILD, J., GODISKA, R., MEAD, D. A., WIGGINS, A. G., GOODMAN, R. M., SZYBALSKI, W. & LILES, M. R. 2011. Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. *Gene,* 475**,** 57-62.

KÄLL, L., KROGH, A. & SONNHAMMER, E. L. 2004. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology,* 338**,** 1027-1036.

KANE, J. F. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Current Opinion in Biotechnology,* 6**,** 494-500.

KANEHISA, M., GOTO, S., HATTORI, M., AOKI-KINOSHITA, K. F., ITOH, M., KAWASHIMA, S., KATAYAMA, T., ARAKI, M. & HIRAKAWA, M. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research,* 34**,** D354-D357.

KANNAN, Y., KOGA, Y., INOUE, Y., HARUKI, M., TAKAGI, M., IMANAKA, T., MORIKAWA, M. & KANAYA, S. 2001. Active Subtilisin-Like Protease from a hyperthermophilic archaeon in a form with a putative prosequence. *Applied and Environmental Microbiology,* 67**,** 2445-2452.

KAPUST, R. & WAUGH, D. 1999. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Science,* 8**,** 1668 - 1674.

KATZEN, F. 2007. Gateway® recombinational cloning: a biological operating system. *Expert Opinion on Drug Discovery,* 2**,** 571-589.

KENNEDY, J., FLEMER, B., JACKSON, S. A., LEJON, D. P. H., P, M. J., O'GARA, F. & DOBSON, D. W. 2010. Marine metagenomics: New tools for the study and exploitation of marine microbial metabolism. *Marine Drugs,* 8**,** 608-628.

KERRIGAN, J., RAGUNATH, C., KANDRA, L., GYÉMÁNT, G., LIPTÁK, A., JÁNOSSY, L., KAPLAN, J. & RAMASUBBU, N. 2008. Modeling and biochemical analysis of the activity of antibiofilm agent Dispersin B. *Acta Biologica Hungarica,* 59**,** 439-451.

KHALIL, A. S. & COLLINS, J. J. 2010. Synthetic biology: applications come of age. *Nature Reviews Genetics,* 11**,** 367-379.

KHAN, A. R. & JAMES, M. N. 1998. Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes. *Protein Sci,* 7**,** 815-836.

KIM, E.-Y., OH, K.-H., LEE, M.-H., KANG, C.-H., OH, T.-K. & YOON, J.-H. 2009. Novel cold-adapted alkaline lipase from an intertidal flat metagenome and proposal for a new family of bacterial lipases. *Applied Environmental Microbiology,* 75**,** 257-260.

KIM, J.-S., KLUSKENS, L. D., DE VOS, W. M., HUBER, R. & VAN DER OOST, J. 2004. Crystal Structure of fervidolysin from *Fervidobacterium pennivorans*, a keratinolytic enzyme related to subtilisin. *Journal of Molecular Biology,* 335**,** 787-797.

KIMHI, Y. & MAGASANIK, B. 1970. Genetic basis of histidine degradation in *Bacillus subtilis*. *Journal of Biological Chemistry,* 245**,** 3545-3548.

KINGSTON, B. & BRENT, R. 2001. Protein Expression. *Current Protocols in Molecular Biology,* 78**,** 16.0.1-16.0.5.

KOBAYASHI, H., UTSUNOMIYA, H., YAMANAKA, H., SEI, Y., KATUNUMA, N., OKAMOTO, K. & TSUGE, H. 2009. Structural basis for the kexin-like serine protease from *Aeromonas sobria* as sepsis-causing factor. *Journal of Biological Chemistry,* 284**,** 27655-27663.

KOH, S., KIM, J., SHIN, H.-J., LEE, D., BAE, J., KIM, D. & LEE, D.-S. 2003. Mechanistic study of the intramolecular conversion of maltose to trehalose by *Thermus caldophilus* GK24 trehalose synthase. *Carbohydrate Research,* 338**,** 1339-1343.

KROGH, A., BROWN, M., MIAN, I. S., SJOLANDER, K. & HAUSSLER, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology,* 235**,** 1501-1531.

KUDLA, G., MURRAY, A. W., TOLLERVEY, D. & PLOTKIN, J. B. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science,* 324**,** 255-258.

KUHN, E., BELLICANTA, G. S. & PELLIZARI, V. H. 2009. New *alk* genes detected in Antarctic marine sediments. *Environmental Microbiology,* 11**,** 669-673.

KULAKOVA, L., GALKIN, A., KURIHARA, T., YOSHIMURA, T. & ESAKI, N. 1999. Cold-Active Serine alkaline protease from the psychrotrophic bacterium *Shewanella* Strain Ac10: Gene Cloning and Enzyme Purification and Characterization. *Applied Environmental Microbiology,* 65**,** 611-617.

KUNIN, V., COPELAND, A., LAPIDUS, A., MAVROMATIS, K. & HUGENHOLTZ, P. 2008. A Bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Review,* 72**,** 557 - 578.

KWON, K., HASSEMAN, J., LATHAM, S., GROSE, C., DO, Y., FLEISCHMANN, R., PIEPER, R. & PETERSON, S. 2011. Recombinant expression and functional analysis of proteases from *Streptococcus pneumoniae, Bacillus anthracis, and Yersinia pestis*. *BMC Biochemistry,* 12**,** 17.

KYTE, J. & DOOLITTLE, R. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology,* 157**,** 105 - 132.

LAMMLE, K., ZIPPER, H., BREUER, M., HAUER, B., BUTA, C., BRUNNER, H. & RUPP, S. 2007. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *Journal of Biotechnology,* 127**,** 575 - 592.

LÄMMLE, K., ZIPPER, H., BREUER, M., HAUER, B., BUTA, C., BRUNNER, H. & RUPP, S. 2007. Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *Journal of Biotechnology,* 127**,** 575-592.

LANGKLOTZ, S., SCHÄKERMANN, M. & NARBERHAUS, F. 2011. Control of lipopolysaccharide biosynthesis by FtsH-mediated proteolysis of LpxC is conserved in enterobacteria but not in all gram-negative bacteria. *Journal of Bacteriology,* 193**,** 1090-1097.

LANOIL, B., SKIDMORE, M., PRISCU, J. C., HAN, S., FOO, W., VOGEL, S. W., TULACZYK, S. & ENGELHARDT, H. 2009. Bacteria beneath the West Antarctic ice sheet. *Environmental Microbiology,* 11**,** 609-615.

LASSY, R. A. L. & MILLER, C. G. 2000. Peptidase E, a peptidase specific for N-terminal aspartic dipeptides, is a serine hydrolase. *Journal of Bacteriology,* 182**,** 2536-2543.

LAURO, F. M., DEMAERE, M. Z., YAU, S., BROWN, M. V., NG, C., WILKINS, D., RAFTERY, M. J., GIBSON, J. A. E., ANDREWS-PFANNKOCH, C., LEWIS, M., HOFFMAN, J. M., THOMAS, T. & CAVICCHIOLI, R. 2011. An integrative study of a meromictic lake ecosystem in Antarctica. *The ISME Journal,* 5**,** 879-895.

LAVALLIE, E., DIBLASIO, E., KOVACIC, S., GRANT, K., SCHENDEL, P. & MCCOY, J. 1993. A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm. *Bio Technology,* 11**,** 187 - 193.

LAYBOURN-PARRY, J. & PEARCE, D. A. 2007. The biodiversity and ecology of Antarctic lakes: models for evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 362**,** 2273-2289.

LAYBOURNE-PARRY, J. & MARCHANT, H. J. 1992. The microbial plankton of freshwater lakes in the Vestfold Hills, Antarctica. *Polar Biology,* 12**,** 405-410.

LEE, B.-M., PARK, Y.-J., PARK, D.-S., KANG, H.-W., KIM, J.-G., SONG, E.-S., PARK, I.-C., YOON, U.-H., HAHN, J.-H., KOO, B.-S., LEE, G.-B., KIM, H., PARK, H.-S., YOON, K.-O., KIM, J.-H., JUNG, C.-H., KOH, N.-H., SEO, J.-S. & GO, S.-J. 2005a. The genome sequence of *Xanthomonas oryzae* pathovar *oryzae* KACC10331, the bacterial blight pathogen of rice. *Nucleic Acids Research,* 33**,** 577-586.

LEE, C., KIBBLEWHITE-ACCINELLI, R., WAGSCHAL, K., ROBERTSON, G. & WONG, D. 2006. Cloning and characterization of a cold-active xylanase enzyme from an environmental DNA library. *Extremophiles,* 10**,** 295 - 300.

LEE, J.-H., LEE, K.-H., KIM, C.-G., LEE, S.-Y., KIM, G.-J., PARK, Y.-H. & CHUNG, S.-O. 2005b. Cloning and expression of a trehalose synthase from *Pseudomonas stutzeri* CJ38 in *Escherichia coli* for the production of trehalose. *Applied Microbiology and Biotechnology,* 68**,** 213-219.

LI, J., SHI, C., GAO, Y., WU, K., SHI, P., LAI, C., CHEN, L., WU, F. & TIAN, C. 2012. Structural Studies of *Mycobacterium tuberculosis* Rv0899 reveal a monomeric membrane-anchoring protein with two separate domains. *Journal of Molecular Biology,* 415**,** 382-392.

LI, L.-L., MCCORKLE, S., MONCHY, S., TAGHAVI, S. & VAN DER LELIE, D. 2009. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnology for Biofuels,* 2**,** 10.

LI, Y., HU, Z., JORDAN, F. & INOUYE, M. 1995. Functional analysis of the propeptide of subtilisin E as an intramolecular chaperone for protein folding : Refolding and inhibitory abilities of propeptide mutants. *Journal of Biological Chemistry,* 270**,** 25127-25132.

LITTHAUER, D., ABBAI, N. S., PIATER, L. A. & VAN HEERDEN, E. 2010. Pitfalls using tributyrin agar screening to detect lipolytic activity in metagenomic studies. *African Journal of Biotechnology,* 9**,** 4282-4285.

LIU, Y. N., TANG, J. L., CLARKE, B. R., DOW, J. M. & DANIELS, M. J. 1990. A multipurpose broad host range cloning vector and its use to characterise an extracellular protease gene of *Xanthomonas campestris* pathovar *campestris*. *Molecular & general genetics : MGG,* 220**,** 433-440.

LOESSNER, M., SCHMELCHER, M., GRALLERT, H. & BRETFELD, F. 2010. Artificial peptidoglycan lysing enzymes and peptidoglycan binding proteins. PCT/EP2009/060716.

LOGARES, R., HAVERKAMP, T. H. A., KUMAR, S., LANZÉN, A., NEDERBRAGT, A. J., QUINCE, C. & KAUSERUD, H. 2012. Environmental microbiology through the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches. *Journal of Microbiological Methods,* 91**,** 106-113.

LOHAN, D. & JOHNSTON, S. 2005. UNU-IAS Report: Bioprospecting in Antarctica.

LONHIENNE, T., GERDAY, C. & FELLER, G. 2000. Psychrophilic enzymes: revisiting the thermodynamic parameters of activation may explain local flexibility. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology,* 1543**,** 1-10.

LORENZ, P., LIEBETON, K., NIEHAUS, F. & ECK, J. 2002. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. *Current Opinion in Biotechnology,* 13**,** 572-577.

LOVELAND-CURTZE, J., MITEVA, V. & BRENCHLEY, J. 2010. Novel ultramicrobacterial isolates from a deep Greenland ice core represent a proposed new species, Chryseobacterium greenlandense sp. nov. *Extremophiles,* 14**,** 61-69.

LUAN, C.-H., QIU, S., FINLEY, J. B., CARSON, M., GRAY, R. J., HUANG, W., JOHNSON, D., TSAO, J., REBOUL, J., VAGLIO, P., HILL, D. E., VIDAL, M., DELUCAS, L. J. & LUO, M. 2004. High-throughput expression of *C. elegans* Proteins. *Genome Research,* 14**,** 2102-2110.

MACGREGOR, E. A., JANEČEK, Š. & SVENSSON, B. 2001. Relationship of sequence and structure to specificity in the α-amylase family of enzymes. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology,* 1546**,** 1-20.

MADHAVAN, V., BHATT, F. & JEFFERY, C. 2010. Recombinant expression screening of *P. aeruginosa* bacterial inner membrane proteins. *BMC Biotechnology,* 10**,** 83.

MALINIČOVÁ, L., PIKNOVÁ, M., PRISTAŠ, P. & JAVORSKÝ, P. 2010. Peptidoglycan hydrolases as novel tool for anti-enterococcal therapy. *Current Research,*

Technology and Education Topics in Applied Microbiology and Microbial Biotechnology. The Formatex Microbiology Book Series. Volume.

MALMSTROM, R. R., KIENE, R. P., COTTRELL, M. T. & KIRCHMAN, D. L. 2004. Contribution of SAR11 bacteria to dissolved dimethylsulfoniopropionate and amino acid uptake in the North Atlantic ocean. *Applied and Environmental Microbiology,* 70**,** 4129-4135.

MARCHLER-BAUER, A., ANDERSON, J., CHITSAZ, F., DERBYSHIRE, M., DEWEESE-SCOTT, C., FONG, J., GEER, L., GEER, R., GONZALES, N., GWADZ, M., HE, S., HURWITZ, D., JACKSON, J., KE, Z., LANCZYCKI, C., LIEBERT, C., LIU, C., LU, F., LU, S., MARCHLER, G., MULLOKANDOV, M., SONG, J., TASNEEM, A., THANKI, N., YAMASHITA, R., ZHANG, D., ZHANG, N. & BRYANT, S. 2009. CDD: Specific functional annotation with the Conserved Domain Database. *Nucleic Acids Research,* 37**,** D205 - D210.

MARGESIN, R. & FELLER, G. 2010. Biotechnological applications of psychrophiles. *Environmental Technology,* 31**,** 835-844.

MARGESIN, R. & MITEVA, V. 2011. Diversity and ecology of psychrophilic microorganisms. *Research in Microbiology,* 162**,** 346-361.

MARGESIN, R., NEUNER, G. & STOREY, K. 2007. Cold-loving microbes, plants, and animals—fundamental and applied aspects. *Naturwissenschaften,* 94**,** 77-99.

MARTINEZ, J., SMITH, D. C., STEWARD, G. F. & AZAM, F. 1996. Variability in ectohydrolytic enzyme activities of pelagic marine bacteria and its significance for substrate processing in the sea. *Aquatic Microbial Ecology,* 10**,** 223-230.

MCBRIDE, M. J. & ZHU, Y. 2013. Gliding motility and Por secretion system genes are widespread among members of the phylum *Bacteroidetes*. *Journal of Bacteriology,* 195**,** 270-278.

MCGINN, S. & GUT, I. G. 2013. DNA sequencing – spanning the generations. *New Biotechnology,* 30**,** 366-372.

MERGULHAO, F., SUMMERS, D. & MONTEIRO, G. 2005. Recombinant protein secretion in *Escherichia coli*. *Biotechnology Advances,* 23**,** 177 - 202.

METZKER, M. L. 2010. Sequencing technologies - the next generation. *Nature Reviews Genetics,* 11**,** 31-46.

MITCHELL, B. G., BRODY, E. A., HOLM-HANSEN, O., MCCLAIN, C. & BISHOP, J. 1991. Light limitation of phytoplankton biomass and macronutrient utilization in the Southern Ocean. *Limnology and oceanography,* 36**,** 1662-1677.

MITCHELL, D. A., MARSHALL, T. K. & DESCHENES, R. J. 1993. Vectors for the inducible overexpression of glutathione S-transferase fusion proteins in yeast. *Yeast,* 9**,** 715-722.

MIYAKE, R., KAWAMOTO, J., WEI, Y.-L., KITAGAWA, M., KATO, I., KURIHARA, T. & ESAKI, N. 2007. Construction of a low-temperature protein expression system using a cold-adapted bacterium, *Shewanella* sp. Strain Ac10, as the Host. *Applied Environmental Microbiology,* 73**,** 4849-4856.

MORITA, R. Y. 1975. Psychrophilic bacteria. *Bacteriological Reviews,* 39**,** 144-167.

MORRIS, L. S., EVANS, J. & MARCHESI, J. R. 2012. A robust plate assay for detection of extracellular microbial protease activity in metagenomic screens and pure cultures. *Journal of Microbiological Methods,* 91**,** 144-146.

MOYER, C. L. & MORITA, R. Y. 2001. Psychrophiles and Psychrotrophs. *eLS.* John Wiley & Sons, Ltd.

NALLAMSETTY, S. & WAUGH, D. S. 2006. Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. *Protein Expression and Purification,* 45**,** 175-182.

NEURATH, H. & WALSH, K. A. 1976. Role of proteolytic enzymes in biological regulation *Proceedings of the National Academy of Sciences,* 73**,** 3825-3832.

NEWCOMB, J., CARLSON, R. & ALDRICH, S. 2007. Genome synthesis and design futures: Implications for the U.S. economy. *Bio Economics Research Associates.*

NG, C., DEMAERE, M. Z., WILLIAMS, T. J., LAURO, F. M., RAFTERY, M., GIBSON, J. A. E., ANDREWS-PFANNKOCH, C., LEWIS, M., HOFFMAN, J. M., THOMAS, T. & CAVICCHIOLI, R. 2010. Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *The ISME Journal*.

NOGUCHI, H., PARK, J. & TAKAGI, T. 2006. MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research.,* 34**,** 5623-5630.

NONAKA, T., FUJIHASHI, M., KITA, A., SAEKI, K., ITO, S., HORIKOSHI, K. & MIKI, K. 2004. The crystal structure of an oxidatively stable subtilisin-like alkaline serine protease, KP-43, with a C-terminal β-barrel domain. *Journal of Biological Chemistry,* 279**,** 47344-47351.

OBAYASHI, Y. & SUZUKI, S. 2005. Proteolytic enzymes in coastal surface seawater: significant activity of endopeptidases and exopeptidases. *Limnology and oceanography***,** 722-726.

OBAYASHI, Y. & SUZUKI, S. 2008. Occurrence of exo-and endopeptidases in dissolved and particulate fractions of coastal seawater. *Aquatic Microbial Ecology,* 50**,** 231-237.

OHOL, Y. M., GOETZ, D. H., CHAN, K., SHILOH, M. U., CRAIK, C. S. & COX, J. S. 2010. *Mycobacterium tuberculosis* MycP1 protease plays a dual role in regulation of ESX-1 secretion and virulence. *Cell Host & Microbe,* 7**,** 210-220.

ONO, A., MIYAZAKI, R., SOTA, M., OHTSUBO, Y., NAGATA, Y. & TSUDA, M. 2007. Isolation and characterization of naphthalene-catabolic genes and plasmids from oil-contaminated soil by using two cultivation-independent approaches. *Applied Microbiology and Biotechnology,* 74**,** 501-510.

OOMEN, C. J., VAN ULSEN, P., VAN GELDER, P., FEIJEN, M., TOMMASSEN, J. & GROS, P. 2004. Structure of the translocator domain of a bacterial autotransporter. *EMBO Journal,* 23**,** 1257-1266.

OTTMANN, C., ROSE, R., HUTTENLOCHER, F., CEDZICH, A., HAUSKE, P., KAISER, M., HUBER, R. & SCHALLER, A. 2009. Structural basis for Ca^{2+} independence and activation by homodimerization of tomato subtilase 3. *Proceedings of the National Academy of Sciences,* 106**,** 17223-17228.

PACE, N., STAHL, D., LANE, D. & OLSEN, G. 1986. The analysis of natural microbial populations by ribosomal RNA sequences. *Advance in Microbial Ecology,* 9**,** 1 - 55.

PAN, Y. T., KOROTH EDAVANA, V., JOURDIAN, W. J., EDMONDSON, R., CARROLL, J. D., PASTUSZAK, I. & ELBEIN, A. D. 2004. Trehalose synthase of *Mycobacterium smegmatis*. *European Journal of Biochemistry,* 271**,** 4259-4269.

PANDEY, A., BENJAMIN, S., SOCCOL, C. R., NIGAM, P., KRIEGER, N. & SOCCOL, V. T. 1999. The realm of microbial lipases in biotechnology. *Biotechnology and Applied Biochemistry,* 29**,** 119-131.

PANDEY, K., SHUKLA, S., SHUKLA, P., GIRI, D., SINGH, J., SINGH, P. & KASHYAP, A. 2004. Cyanobacteria in Antarctica: Ecology, physiology and cold adaptation. *Cellular and Molecular Biology,* 50**,** 575-584.

PARACHIN, N. & GORWA-GRAUSLUND, M. 2011. Isolation of xylose isomerases by sequence- and function-based screening from a soil metagenomic library. *Biotechnology for Biofuels,* 4**,** 9.

PARK, H.-J., JEON, J. H., KANG, S. G., LEE, J.-H., LEE, S.-A. & KIM, H.-K. 2007. Functional expression and refolding of new alkaline esterase, EM2L8 from deep-sea sediment metagenome. *Protein Expression and Purification,* 52**,** 340-347.

PARKS, D. H. & BEIKO, R. G. 2010. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics,* 26**,** 715-721.

PARSONS, L. M., LIN, F. & ORBAN, J. 2006. Peptidoglycan recognition by Pal, an Outer Membrane Lipoprotein. *Biochemistry,* 45**,** 2122-2128.

PELLET, J., TAFFOREAU, L., LUCAS-HOURANI, M., NAVRATIL, V., MEYNIEL, L., ACHAZ, G., GUIRONNET-PAQUET, A., AUBLIN-GEX, A., CAIGNARD, G. & CASSONNET, P. 2010. ViralORFeome: An integrated database to generate a versatile collection of viral ORFs. *Nucleic Acids Research,* 38**,** D371-D378.

PELLETIER, E., KREIMEYER, A., BOCS, S., ROUY, Z., GYAPAY, G., CHOUARI, R., RIVIÈRE, D., GANESAN, A., DAEGELEN, P., SGHIR, A., COHEN, G. N., MÉDIGUE, C., WEISSENBACH, J. & LE PASLIER, D. 2008. *Candidatus* Cloacamonas Acidaminovorans: Genome sequence reconstruction provides a first glimpse of a new bacterial division. *Journal of Bacteriology,* 190**,** 2572-2579.

PENG, Y., LEUNG, H. C. M., YIU, S. M. & CHIN, F. Y. L. 2011. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics,* 27**,** 94-101.

PETERSEN, T. N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods,* 8.

PIETTE, F., STRUVAY, C., GODIN, A., CIPOLLA, A. & FELLER, G. 2012. Life in the Cold: Proteomics of the Antarctic Bacterium *Pseudoalteromonas haloplanktis*. *In:* HEAZLEWOOD, J. (ed.) *Proteomic Applications in Biology.* InTech.

PINHASSI, J., SALA, M. M., HAVSKUM, H., PETERS, F., GUADAYOL, Ò., MALITS, A. & MARRASÉ, C. 2004. Changes in bacterioplankton composition under different phytoplankton regimens. *Applied and Environmental Microbiology,* 70**,** 6753-6766.

PLOTKIN, J. B., KUDLA, G. & TAFFOREAU, L. 2011. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics,* 12**,** 32-42.

POLGAR, L. 2002. The prolyl oligopeptidase family. *Cellular and Molecular Life Sciences CMLS,* 59**,** 349-362.

POPHAM, D. L. & YOUNG, K. D. 2003. Role of penicillin-binding proteins in bacterial cell morphogenesis. *Current Opinion in Microbiology,* 6**,** 594-599.

PORETSKY, R. S., SUN, S., MOU, X. & MORAN, M. A. 2010. Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environmental Microbiology,* 12**,** 616-627.

POWELL, L., BOWMAN, J., SKERRATT, J., FRANZMANN, P. & BURTON, H. 2005. Ecology of a novel Synechococcus clade occurring in dense populations in saline Antarctic lakes. *Marine Ecology Progress Series,* 291**,** 65-80.

PRICE, P. B. 2007. Microbial life in glacial ice and implications for a cold origin of life. *FEMS Microbiology Ecology,* 59**,** 217-231.

PRISCU, J. C. & CHRISTENER, B. C. 2004. *Earth's biosphere,* Washington DC, ASM Press.

PRISCU, J. C., CHRISTNERB.C., FOREMANC.F. & G., R.-B. 2007. Biological material in ice cores. *In:* ELIAS, S. A. (ed.) *Encyclopedia of Quaternary Science.* UK: Elsevier.

PRISCU, J. C., FRITSEN, C. H., ADAMS, E. E., GIOVANNONI, S. J., PAERL, H. W., MCKAY, C. P., DORAN, P. T., GORDON, D. A., LANOIL, B. D. & PINCKNEY, J. L. 1998. Perennial Antarctic Lake Ice: An Oasis for Life in a Polar Desert. *Science,* 280**,** 2095-2098.

PULIDO, M., SAITO, K., TANAKA, S.-I., KOGA, Y., MORIKAWA, M., TAKANO, K. & KANAYA, S. 2006. Ca^{2+}-dependent maturation of subtilisin from a hyperthermophilic archaeon, *Thermococcus kodakaraensis*: the propeptide is a potent inhibitor of the mature domain but is not required for its folding. *Applied and Environmental Microbiology,* 72**,** 4154-4162.

RAJAGOPALA, S., YAMAMOTO, N., ZWEIFEL, A., NAKAMICHI, T., HUANG, H.-K., MENDEZ-RIOS, J., FRANCA-KOH, J., BOORGULA, M., FUJITA, K., SUZUKI, K.-I., HU, J., WANNER, B., MORI, H. & UETZ, P. 2010. The *Escherichia coli* K-12 ORFeome: a resource for comparative molecular microbiology. *BMC Genomics,* 11**,** 470.

RANKIN, L. M., GIBSON, J. A. E., FRANZMANN, P. D. & BURTON, H. R. 1999. The chemical stratification and microbial communities of Ace Lake, Antarctica: A review of the characteristics of a marine-derived meromictic lake. *Polarforschung,* 66**,** 33-52.

RAO, M. B., TANKSALE, A. M., GHATGE, M. S. & DESHPANDE, V. V. 1998. Molecular and biotechnological aspects of microbial proteases. *Microbiology and Molecular Biology Reviews,* 62**,** 597-635.

RATH, J., WU, K. Y., HERNDL, G. J. & DELONG, E. F. 1998. High phylogenetic diversity in a marine-snow-associated bacterial assemblage. *Aquatic Microbial Ecology,* 14**,** 261-269.

RAWLINGS, N. D. & BARRETT, A. J. 1993. Evolutionary families of peptidases. *Biochemical Journal,* 290**,** 205-218.
RAWLINGS, N. D., BARRETT, A. J. & BATEMAN, A. 2012. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research,* 40**,** D343-D350.
RAWLINGS, N. D., MORTON, F. R. & BARRETT, A. J. 2006. MEROPS:The peptidase database. *Nucleic Acids Research.,* 34**,** D270-272.
REDDY, N., NIMMAGADDA, A. & RAO, K. S. 2004. An overview of the microbial α-amylase family. *African Journal of Biotechnology,* 2**,** 645-648.
RIBITSCH, D., HEUMANN, S., KARL, W., GERLACH, J., LEBER, R., BIRNER-GRUENBERGER, R., GRUBER, K., EITELJOERG, I., REMLER, P., SIEGERT, P., LANGE, J., MAURER, K. H., BERG, G., GUEBITZ, G. M. & SCHWAB, H. 2012. Extracellular serine proteases from Stenotrophomonas maltophilia: Screening, isolation and heterologous expression in E. coli. *Journal of Biotechnology,* 157**,** 140-147.
RO, D.-K., PARADISE, E. M., OUELLET, M., FISHER, K. J., NEWMAN, K. L., NDUNGU, J. M., HO, K. A., EACHUS, R. A., HAM, T. S., KIRBY, J., CHANG, M. C. Y., WITHERS, S. T., SHIBA, Y., SARPONG, R. & KEASLING, J. D. 2006. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature,* 440**,** 940.
ROH, C. & VILLATTE, F. 2008. Isolation of a low-temperature adapted lipolytic enzyme from uncultivated micro-organism. *Journal of Applied Microbiology,* 105**,** 116-123.
ROY, A., KUCUKURAL, A. & ZHANG, Y. 2010. I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols,* 5**,** 725-738.
RUAL, J.-F., HILL, D. E. & VIDAL, M. 2004. ORFeome projects: gateway between genomics and omics. *Current Opinion in Chemical Biology,* 8**,** 20-25.
RUDOLPH, R. & LILIE, H. 1996. In vitro folding of inclusion body proteins. *The FASEB Journal,* 10**,** 49-56.
RUSCH, D. B., HALPERN, A. L., SUTTON, G., HEIDELBERG, K. B., WILLIAMSON, S., YOOSEPH, S., WU, D., EISEN, J. A., HOFFMAN, J. M., REMINGTON, K., BEESON, K., TRAN, B., SMITH, H., BADEN-TILLSON, H., STEWART, C., THORPE, J., FREEMAN, J., ANDREWS-PFANNKOCH, C., VENTER, J. E., LI, K., KRAVITZ, S., HEIDELBERG, J. F., UTTERBACK, T., ROGERS, Y.-H., FALCÃ^{3}N, L. I., SOUZA, V., BONILLA-ROSSO, G. N., EGUIARTE, L. E., KARL, D. M., SATHYENDRANATH, S., PLATT, T., BERMINGHAM, E., GALLARDO, V., TAMAYO-CASTILLO, G., FERRARI, M. R., STRAUSBERG, R. L., NEALSON, K., FRIEDMAN, R., FRAZIER, M. & VENTER, J. C. 2007. The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLOS Biology,* 5**,** e77.
RUSSELL, N. J. 2003. Psychrophily and resistance to low temperatures. *Extremophiles (Life under extreme environmental Condition), Encyclopedia of Life Support Systems (EOLSS) Developed under the Auspices of the UNESCO. Eolss Publishers, Oxford, UK. http://www. eolss. net.*
SAKAGUCHI, M., MATSUZAKI, M., NIIMIYA, K., SEINO, J., SUGAHARA, Y. & KAWAKITA, M. 2007. Role of proline residues in conferring thermostability on aqualysin I. *Journal of Biochemistry,* 141**,** 213-220.
SAKOH, M., ITO, K. & AKIYAMA, Y. 2005. Proteolytic activity of HtpX, a membrane-bound and stress-controlled protease from *Escherichia coli*. *Journal of Biological Chemistry,* 280**,** 33305-33310.
SALCHER, M. M., JAKOB PERNTHALER & POSCH, T. 2010. Spatiotemporal distribution and activity patterns of bacteria from three phylogenetic groups in an oligomesotrophic lake. *Limnology and Oceanography,* 55**,** 846-856.
SAUVAGE, E., KERFF, F., TERRAK, M., AYALA, J. A. & CHARLIER, P. 2008. The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. *FEMS Microbiology Reviews,* 32**,** 234-258.
SCHATZ, M. C., PHILLIPPY, A. M., SHNEIDERMAN, B. & SALZBERG, S. L. 2007. Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome biology,* 8**,** R34.

SCHEIN, C. H. 1989. Production of soluble recombinant proteins in bacteria. *Nature Biotechnology,* 7**,** 1141-1149.

SCHLÜTER, A., BEKEL, T., DIAZ, N. N., DONDRUP, M., EICHENLAUB, R., GARTEMANN, K.-H., KRAHN, I., KRAUSE, L., KRÖMEKE, H., KRUSE, O., MUSSGNUG, J. H., NEUWEGER, H., NIEHAUS, K., PÜHLER, A., RUNTE, K. J., SZCZEPANOWSKI, R., TAUCH, A., TILKER, A., VIEHÖVER, P. & GOESMANN, A. 2008. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *Journal of Biotechnology,* 136**,** 77-90.

SCHMIDT, T., DELONG, E. & PACE, N. 1991a. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology,* 173**,** 4371 - 4378.

SCHOLZ, M. B., LO, C.-C. & CHAIN, P. S. G. 2012. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology,* 23**,** 9-15.

SEOW, K. T., MEURER, G., GERLITZ, M., WENDT-PIENKOWSKI, E., HUTCHINSON, C. R. & DAVIES, J. 1997. A study of iterative type II polyketide synthases, using bacterial genes cloned from soil DNA: A means to access and use genes from uncultured microorganisms. *Journal of Bacteriology,* 179**,** 7360-7368.

SHAH, H. N. & WILLIAMS, R. A. D. 1987. Catabolism of aspartate and asparagine by *Bacteroides intermedius* and *Bacteroides gingivalis*. *Current Microbiology,* 15**,** 313-318.

SHARP, P. M. & LI, W.-H. 1987. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research,* 15**,** 1281-1295.

SHEEHAN, S. M. & SWITZER, R. L. 1990. Intracellular serine protease 1 of *Bacillus subtilis* is formed in vivo as an unprocessed, active protease in stationary cells. *Journal of Bacteriology,* 172**,** 473-476.

SHIH, Y.-P., KUNG, W.-M., CHEN, J.-C., YEH, C.-H., WANG, A. H. J. & WANG, T.-F. 2002. High-throughput screening of soluble recombinant proteins. *Protein Science,* 11**,** 1714-1719.

SHIMAMOTO, S., MORIYAMA, R., SUGIMOTO, K., MIYATA, S. & MAKINO, S. 2001. Partial characterization of an enzyme fraction with protease activity which converts the spore peptidoglycan hydrolase (SleC) precursor to an active enzyme during germination of *Clostridium perfringens* S40 spores and analysis of a gene cluster involved in the activity. *Journal of Bacteriology,* 183**,** 3742-3751.

SHINDE, U. & INOUYE, M. 1995. Folding pathway mediated by an intramolecular chaperone: Characterization of the structural changes in pro-subtilisin E coincident with autoprocessing. *Journal of Molecular Biology,* 252**,** 25-30.

SHINDE, U. & INOUYE, M. 1996. Propeptide-mediated folding in subtilisin: The intramolecular chaperone concept. *Advances in Experimental Medicine and Biology,* 379**,** 147.

SHINDE, U. & INOUYE, M. 2000. Intramolecular chaperones: polypeptide extensions that modulate protein folding. *Seminars in cell & developmental biology,* 11**,** 35-44.

SHOJI, M., SATO, K., YUKITAKE, H., KONDO, Y., NARITA, Y., KADOWAKI, T., NAITO, M. & NAKAYAMA, K. 2011. Por secretion system-dependent secretion and glycosylation of *Porphyromonas gingivalis* hemin-binding protein 35. *PLoS ONE,* 6**,** e21372.

SIDDIQUI, K. & CAVICCHIOLI, R. 2006. Cold-adapted enzymes. *Annual Review of Biochemistry,* 75**,** 403 - 433.

SIEZEN, R., KUIPERS, O. & VOS, W. 1996. Comparison of lantibiotic gene clusters and encoded proteins. *Antonie van Leeuwenhoek,* 69**,** 171-184.

SIEZEN, R. J. & LEUNISSEN, J. A. 1997. Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Science,* 6**,** 501-523.

SIEZEN, R. J., RENCKENS, B. & BOEKHORST, J. 2007. Evolution of prokaryotic subtilases: Genome-wide analysis reveals novel subfamilies with different catalytic residues. *Proteins: Structure, Function, and Bioinformatics,* 67**,** 681-694.

SIMON, C. & DANIEL, R. 2009. Achievements and new knowledge unraveled by metagenomic approaches. *Applied Microbiology and Biotechnology,* 85**,** 265-276.

SIMON, C. & DANIEL, R. 2011. Metagenomic analyses: Past and future trends. *Applied and Environmental Microbiology,* 77**,** 1153-1161.

SIMON, C., HERATH, J., ROCKSTROH, S. & DANIEL, R. 2009a. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomiclibraries derived from glacial ice. *Applied Environmental Microbiology,* 75**,** 2964-2968.

SIMON, C., WIEZER, A., STRITTMATTER, A. W. & DANIEL, R. 2009b. Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Applied and Environmental Microbiology,* 75**,** 7519-7526.

ŠKALAMERA, D., RANALL, M. V., WILSON, B. M., LEO, P., PURDON, A. S., HYDE, C., NOURBAKHSH, E., GRIMMOND, S. M., BARRY, S. C., GABRIELLI, B. & GONDA, T. J. 2011. A high-throughput platform for lentiviral overexpression screening of the human ORFeome. *PLoS ONE,* 6**,** e20057.

SMITH, D. C., SIMON, M., ALLDREDGE, A. L. & AZAM, F. 1992. Intense hydrolytic enzyme activity on marine aggregates and implications for rapid particle dissolution. *Nature,* 359**,** 139-142.

SMITH, G. R. & MAGASANIK, B. 1971. The two operons of the histidine utilization system in Salmonella typhimurium. *Journal of Biological Chemistry,* 246**,** 3330-3341.

SONG, J. M., AN, Y. J., KANG, M. H., LEE, Y.-H. & CHA, S.-S. 2012. Cultivation at 6–10° C is an effective strategy to overcome the insolubility of recombinant proteins in *Escherichia coli*. *Protein Expression and Purification,* 82**,** 297-301.

SORENSEN, H. & MORTENSEN, K. 2005. Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *Journal of Biotechnology,* 115**,** 113 - 128.

SØRHAUG, T. & STEPANIAK, L. 1997. Psychrotrophs and their enzymes in milk and dairy products: Quality aspects. *Trends in Food Science & Technology,* 8**,** 35-41.

SOWELL, S. M., ABRAHAM, P. E., SHAH, M., VERBERKMOES, N. C., SMITH, D. P., BAROFSKY, D. F. & GIOVANNONI, S. J. 2011. Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *The ISME Journal,* 5**,** 856-865.

SOWELL, S. M., WILHELM, L. J., NORBECK, A. D., LIPTON, M. S., NICORA, C. D., BAROFSKY, D. F., CARLSON, C. A., SMITH, R. D. & GIOVANONNI, S. J. 2008. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *The ISME Journal,* 3**,** 93-105.

STAHL, M. L. & FERRARI, E. 1984. Replacement of the *Bacillus subtilis* subtilisin structural gene with an In vitro-derived deletion mutation. *Journal of Bacteriology,* 158**,** 411-418.

STAM, M. R., DANCHIN, E. G. J., RANCUREL, C., COUTINHO, P. M. & HENRISSAT, B. 2006. Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of α-amylase-related proteins. *Protein Engineering Design and Selection,* 19**,** 555-562.

STANLEY, S. A., RAGHAVAN, S., HWANG, W. W. & COX, J. S. 2003. Acute infection and macrophage subversion by *Mycobacterium tuberculosis* require a specialized secretion system. *Proceedings of the National Academy of Sciences,* 100**,** 13001-13006.

STEELE, H. L., JAEGER, K. E., DANIEL, R. & STREIT, W. R. 2009. Advances in recovery of novel biocatalysts from metagenomes. *Journal of Molecular Microbiology and Biotechnology,* 16**,** 25-37.

SUBBIAN, E., YABUTA, Y. & SHINDE, U. P. 2005. Folding pathway mediated by an intramolecular chaperone: Intrinsically unstructured propeptide modulates stochastic activation of subtilisin. *Journal of Molecular Biology,* 347**,** 367-383.

TAKACS, C. D., PRISCU, J. C. & MCKNIGHT, D. 2001. Bacterial dissolved organic carbon demand in McMurdo Dry Valley lakes, Antarctica. *Limnol. Oceanography*, 1189-1194.

TAMURA, K., PETERSON, D., PETERSON, N., STECHER, G., NEI, M. & KUMAR, S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and Maximum Parsimony methods. *Molecular Biology and Evolution,* 28, 2731-2739.

TANAKA, S.-I., TAKEUCHI, Y., MATSUMURA, H., KOGA, Y., TAKANO, K. & KANAYA, S. 2008. Crystal structure of Tk-subtilisin folded without propeptide: Requirement of propeptide for acceleration of folding. *FEBS Letters,* 582, 3875-3878.

TANG, Y., SHIGEMATSU, T., MORIMURA, S. & KIDA, K. 2005. Microbial community analysis of mesophilic anaerobic protein degradation process using bovine serum albumin (BSA)-fed continuous cultivation. *Journal of Bioscience and Bioengineering,* 99, 150-164.

TARAO, M., JEZBERA, J. & HAHN, M. W. 2009. Involvement of cell surface structures in size-independent grazing resistance of freshwater *Actinobacteria*. *Applied and Environmental Microbiology,* 75, 4720-4726.

TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L. & NIKOLSKAYA, A. N. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* 4, 41.

TATUSOV, R. L., KOONIN, E. V. & LIPMAN, D. J. 1997. A genomic perspective on protein families. *Science,* 278, 631-637.

TAUPP, M., MEWIS, K. & HALLAM, S. J. 2011. The art and design of functional metagenomic screens. *Current Opinion in Biotechnology,* 22, 465-472.

TORRES, M., DOLCET, M. M., SALA, N. & CANELA, R. 2003. Endophytic fungi associated with Mediterranean plants as a source of mycelium-bound lipases. *Journal of Agricultural and Food Chemistry,* 51, 3328-3333.

TORSVIK, V., GOKSOYR, J. & DAAE, F. 1990. High diversity in DNA of soil bacteria. *Applied Environmental Microbiology,* 56, 782 - 787.

TRIPATHI, L. P. & SOWDHAMINI, R. 2008. Genome-wide survey of prokaryotic serine proteases: analysis of distribution and domain architectures of five serine protease families in prokaryotes. *BMC Genomics,* 9, 549.

TROESCHEL, S. C., THIES, S., LINK, O., REAL, C. I., KNOPS, K., WILHELM, S., ROSENAU, F. & JAEGER, K.-E. 2012. Novel broad host range shuttle vectors for expression in *Escherichia coli, Bacillus subtilis and Pseudomonas putida. Journal of Biotechnology,* 161, 71-79.

TSUMOTO, K., EJIMA, D., KUMAGAI, I. & ARAKAWA, T. 2003. Practical considerations in refolding proteins from inclusion bodies. *Protein Expression and Purification,* 28, 1-8.

TSUSAKI, K., NISHIMOTO, T., NAKADA, T., KUBOTA, M., CHAEN, H., SUGIMOTO, T. & KURIMOTO, M. 1996. Cloning and sequencing of trehalose synthase gene from *Pimelobacter* sp. R48. *Biochimica et Biophysica Acta (BBA)-General Subjects,* 1290, 1-3.

TUTINO, M., DUILIO, A., PARRILLI, E., REMAUT, E., SANNIA, G. & MARINO, G. 2001. A novel replication element from an Antarctic plasmid as a tool for the expression of proteins at low temperature. *Extremophiles,* 5, 257-264.

TUTINO, M. L., DUILIO, A., MORETTI, M. A., SANNIA, G. & MARINO, G. 2000. A rolling-circle plasmid from *Psychrobacter* sp. TA144: Evidence for a novel rep subfamily. *Biochemical and Biophysical Research Communications,* 274, 488-495.

UCHIYAMA, T., ABE, T., IKEMURA, T. & WATANABE, K. 2005. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nature Biotechnology,* 23, 88.

UCHIYAMA, T. & MIYAZAKI, K. 2009. Functional metagenomics for enzyme discovery: challenges to efficient screening. *Current Opinion in Biotechnology,* 20**,** 616-622.

UCHIYAMA, T. & MIYAZAKI, K. 2010. Product-induced gene expression, a product-responsive reporter assay used to screen metagenomic libraries for enzyme-encoding genes. *Applied and Environmental Microbiology,* 76**,** 7029-7035.

UEDA, M., GOTO, T., NAKAZAWA, M., MIYATAKE, K., SAKAGUCHI, M. & INOUYE, K. 2010. A novel cold-adapted cellulase complex from *Eisenia foetida*: Characterization of a multienzyme complex with carboxymethylcellulase, β-glucosidase, β-1,3 glucanase, and β-xylosidase. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology,* 157**,** 26-32.

VARIN, T., LOVEJOY, C., JUNGBLUT, A. D., VINCENT, W. F. & CORBEIL, J. 2012. Metagenomic analysis of stress genes in microbial mat communities from Antarctica and the High Arctic. *Applied and Environmental Microbiology,* 78**,** 549-559.

VARTOUKIAN, S. R., PALMER, R. M. & WADE, W. G. 2010. Strategies for culture of 'unculturable'bacteria. *FEMS microbiology letters,* 309**,** 1-7.

VASANTHA, N., THOMPSON, L. D., RHODES, C., BANNER, C., NAGLE, J. & FILPULA, D. 1984. Genes for alkaline protease and neutral protease from *Bacillus amyloliquefaciens* contain a large open reading frame between the regions coding for signal sequence and mature protein. *Journal of Bacteriology,* 159**,** 811-819.

VENTER, J. C., REMINGTON, K., HEIDELBERG, J. F., HALPERN, A. L., RUSCH, D., EISEN, J. A., WU, D., PAULSEN, I., NELSON, K. E. & NELSON, W. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science,* 304**,** 66 - 74.

VÉVODOVÁ, J., GAMBLE, M., KÜNZE, G., ARIZA, A., DODSON, E., JONES, D. D. & WILSON, K. S. 2010. Crystal structure of an intracellular subtilisin reveals novel structural features unique to this subtilisin family. *Structure,* 18**,** 744-755.

VISWESWARAN, G. R. R., DIJKSTRA, B. W. & KOK, J. 2011. Murein and pseudomurein cell wall binding domains of bacteria and archaea—a comparative view. *Applied Microbiology and Biotechnology,* 92**,** 921-928.

VOGET, S., STEELE, H. & STREIT, W. R. 2005. Metagenomes - an unlimited resource for novel genes, biocatalysts and metabolites. *Minerva Biotechnologica,* 17.

VON MERING, C., HUGENHOLTZ, P., RAES, J., TRINGE, S., DOERKS, T., JENSEN, L., WARD, N. & BORK, P. 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science,* 315**,** 1126 - 1130.

WAGNER, R., AIGNER, H. & FUNK, C. 2012. FtsH proteases located in the plant chloroplast. *Physiologia Plantarum,* 145**,** 203-214.

WAINWRIGHT, M., WICKRAMASINGHE, N., NARLIKAR, J. & RAJARATNAM, P. 2003. Microorganisms cultured from stratospheric air samples obtained at 41 km. *FEMS Microbiology Letters,* 218**,** 161-165.

WANG, G., LUO, H., WANG, Y., HUANG, H., SHI, P., YANG, P., MENG, K., BAI, Y. & YAO, B. 2011. A novel cold-active xylanase gene from the environmental DNA of goat rumen contents: Direct cloning, expression and enzyme characterization. *Bioresource Technology,* 102**,** 3330-3336.

WANG, J. J., SWAISGOOD, H. E. & SHIH, J. C. 2003. Bioimmobilization of keratinase using *Bacillus subtilis* and *Escherichia coli* systems. *Biotechnology and Bioengineering,* 81**,** 421-429.

WASCHKOWITZ, T., ROCKSTROH, S. & DANIEL, R. 2009. Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. *Applied Environmental Microbiology,* 75**,** 2506-2516.

WAUGH, D. S. 2005. Making the most of affinity tags. *Trends in Biotechnology,* 23**,** 316-320.

WEI, P., BAI, L., SONG, W. & HAO, G. 2009. Characterization of two soil metagenome-derived lipases with high specificity for p-nitrophenyl palmitate. *Archives of Microbiology,* 191**,** 233-240.

WEI, Y.-T., ZHU, Q.-X., LUO, Z.-F., LU, F.-S., CHEN, F.-Z., WANG, Q.-Y., HUANG, K., MENG, J.-Z., WANG, R. & HUANG, R.-B. 2004. Cloning, expression and identification of a new

trehalose synthase gene from *Thermobifida fusca* genome. *Acta Biochimica et Biophysica Sinica,* 36**,** 477-484.

WELCH, M., GOVINDARAJAN, S., NESS, J. E., VILLALOBOS, A., GURNEY, A., MINSHULL, J. & GUSTAFSSON, C. 2009a. Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS ONE,* 4**,** e7002.

WELCH, M., VILLALOBOS, A., GUSTAFSSON, C. & MINSHULL, J. 2009b. You're one in a googol: optimizing genes for protein expression. *Journal of the Royal Society Interface,* 6**,** S467-S476.

WILKINS, D., LAURO, F. M., WILLIAMS, T. J., DEMAERE, M. Z., BROWN, M. V., HOFFMAN, J. M., ANDREWS-PFANNKOCH, C., MCQUAID, J. B., RIDDLE, M. J., RINTOUL, S. R. & CAVICCHIOLI, R. 2013. Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. *Environmental Microbiology,* 15**,** 1318-1333.

WILLIAMS, T., LONG, E., EVANS, F., DEMAERE, M. Z., LAURO, F. M., RAFTERY, M. J., DUCKLOW, H., GRZYMSKI, J. J., MURRAY, A. E. & CAVICCHIOLI, R. 2012. A metaproteomic assessment of winter and summer bacterioplankton from Antarctic Peninsula coastal surface waters. *The ISME Journal,* 6**,** 1883-1900.

WILLIAMS, T. J., WILKINS, D., LONG, E., EVANS, F., DEMAERE, M. Z., RAFTERY, M. J. & CAVICCHIOLI, R. 2013. The role of planktonic *Flavobacteria* in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. *Environmental Microbiology,* 15**,** 1302-1317.

WOLF, A., KRÄMER, R. & MORBACH, S. 2003. Three pathways for trehalose metabolism in *Corynebacterium glutamicum* ATCC13032 and their significance in response to osmotic stress. *Molecular Microbiology,* 49**,** 1119-1134.

WOOD, T. M. & BHAT, K. M. 1988. Methods for measuring cellulase activities. *Methods in Enzymology.,* 160**,** 87-112.

WRIGHT, S. W., VAN DEN ENDEN, R. L., PEARCE, I., DAVIDSON, A. T., SCOTT, F. J. & WESTWOOD, K. J. 2010. Phytoplankton community structure and stocks in the Southern Ocean (30–80°E) determined by CHEMTAX analysis of HPLC pigment signatures. *Deep Sea Research Part II: Topical Studies in Oceanography,* 57**,** 758-778.

XIULI, W., HONGBIAO, D., MING, Y. & YU, Q. 2009. Gene cloning, expression, and characterization of a novel trehalose synthase from *Arthrobacter aurescens*. *Applied Microbiology and Biotechnology,* 83**,** 477-482.

YAU, S., LAURO, F. M., DEMAERE, M. Z., BROWN, M. V., THOMAS, T., RAFTERY, M. J., ANDREWS-PFANNKOCH, C., LEWIS, M., HOFFMAN, J. M., GIBSON, J. A. & CAVICCHIOLI, R. 2011. Virophage control of antarctic algal host–virus dynamics. *Proceedings of the National Academy of Sciences,* 108**,** 6163-6168.

YAU, S., LAURO, F. M., WILLIAMS, T. J., DEMAERE, M. Z., BROWN, M. V., RICH, J., GIBSON, J. A. & CAVICCHIOLI, R. 2013. Metagenomic insights into strategies of carbon conservation and unusual sulfur biogeochemistry in a hypersaline Antarctic lake. *The ISME Journal*.

YERGEAU, E., HOGUES, H., WHYTE, L. G. & GREER, C. W. 2010. The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. *The ISME Journal,* 4**,** 1206-1214.

YIN, J., LI, G., REN, X. & HERRLER, G. 2007. Select what you need: a comparative evaluation of the advantages and limitations of frequently used expression systems for foreign genes. *Journal of Biotechnology,* 127**,** 335-347.

YOOSEPH, S., SUTTON, G., RUSCH, D. B., HALPERN, A. L., WILLIAMSON, S. J., REMINGTON, K., EISEN, J. A., HEIDELBERG, K. B., MANNING, G., LI, W., JAROSZEWSKI, L., CIEPLAK, P., MILLER, C. S., LI, H., MASHIYAMA, S. T., JOACHIMIAK, M. P., VAN BELLE, C., CHANDONIA, J.-M., SOERGEL, D. A., ZHAI, Y., NATARAJAN, K., LEE, S., RAPHAEL, B. J., BAFNA, V., FRIEDMAN, R., BRENNER, S. E., GODZIK, A., EISENBERG, D., DIXON, J. E., TAYLOR, S. S., STRAUSBERG, R. L., FRAZIER, M. & VENTER, J. C.

2007. The *Sorcerer II* Global Ocean Sampling expedition: Expanding the universe of protein families. *PLOS Biology,* 5**,** e16.

YUE, M., WU, X., GONG, W. & DING, H. 2009. Molecular cloning and expression of a novel trehalose synthase gene from *Enterobacter hormaechei*. *Microbial Cell Factories,* 8**,** 34.

YUTIN, N., SUZUKI, M. T., TEELING, H., WEBER, M., VENTER, J. C., RUSCH, D. B. & BÉJÀ, O. 2007. Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. *Environmental Microbiology,* 9**,** 1464-1475.

ZHANG, R., PAN, Y. T., HE, S., LAM, M., BRAYER, G. D., ELBEIN, A. D. & WITHERS, S. G. 2011. Mechanistic analysis of trehalose synthase from *Mycobacterium smegmatis. Journal of Biological Chemistry,* 286**,** 35601-35609.

ZHANG, Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics,* 9**,** 40.

ZWARTZ, D., BIRD, M., STONE, J. & LAMBECK, K. 1998. Holocene sea-level change and ice-sheet history in the Vestfold Hills, East Antarctica. *Earth and Planetary Science Letters,* 155**,** 131-145.

Appendix A

The components of stock buffers and solutions

1. Ammonium persulfate (APS) (10% (w/v)) - 0.1 g APS is dissolved into 800 µL of dd$H_2O$and stored at 4°C.
2. Ampicillin – 100 mg of ampicilin disodium salt is dissolved in 800 µL ddH_2O. The volume was adjusted to 1mL.
3. $CaCl_2$, 1M - the buffer was prepared with 147 g $CaCl_2$.2H2O in 500 mL of ddH_2O. The volume was adjusted to 1L.
4. CHAPS (100mM) – 6.149 g of CHAPS is dissolved in 100 ml of water.
5. Chloramphenicol - 0.34 g of chloramphenicol is dissolved into 10 ml of 100% ethanol. The solution was filtered through a 0.22 µm filter to sterilize, aliquoted and stored at –20°C. It was used at 1:1000 dilutions in LB or LB Agar.
6. Dinitrosalicylic Acid Reagent (DNS) - 10g Dinitrosalicyclic, 0.5 g Sodium Sulfite acid, and 800 g Rochelle salt are dissolved in 500mL of 2% NaOH solution then diluted to 1L with ddH_2O.
7. DTT, 1M - 15.45g of DTT is dissolved in 100 mL of ddH_2O. The solution was aliquot into 2 mL tubes and stored at –20°C.
8. EDTA, 0.5 M (pH8) - 186.1 g of Na_2EDTA.2H_2O is dissolved in 800mL ddH_2O. The pH was adjusted to 8 with NaOH (about 20g pellets) and volume to 1L.
9. Ethidium Bromide - 0.2g of ethidium bromide is mixed into 20 mL of ddH_2O and stored at 4°C in the dark (stock solution is at 10 mg/mL).
10. HEPES buffer 1M -238.3 g of HEPES is dissolved in 800 mL of ddH_2O. NaOH was used to adjust the pH to 7.5 before adding ddH_2O to 1L.
11. Imidazole 1M - 68.08 g of imidazole is dissolved in 500ml ddH_2O and the volume was adjusted to 1L.
12. IPTG (MW= 283.3) - 2g IPTG is dissolved in 8 mL ddH_2O adjusted to 10mL. The solution is filter sterilized with 0.22µm filter and stored as 1mL aliquots at –20°C.
13. Kanamycin – 10 mg of kanamycin is dissolved into 800 µL ddH_2O. The volume was adjusted to 1 mL and stored at -20°C.
14. LB broth-10 g bactotryptone, 5 g yeast extract, 10 g NaCl, per L; pH 7. Add 15 g agar per L for LB agar.
15. $MgCl_2$ - by dissolving 203.30 g $MgCl_2$.6H_2O in 800 mL ddH_2O before adjusting the volume to 1 L.

16. MUF-butyrate stock solution - Prepared at a concentration of 25 mM in ethyleneglycol monomethylether stored at −20°C. The final concentration used was 100 µM.
17. Native PAGE gel – Stacking gel (4%)- the gel was prepared with 1 mL of 40% acrylamide/bis solution (37.5:1 crosslinker ratio), 2.5 mL of 0.5M of Tris (pH6.8), 6.120 mL of ddH_2O, 10µL of TEMED and 330 µL of 0.05% (w/v) APS. Resolving gel (10%) the gel was prepared with 2.5 mL of 40% acrylamide/bis solution (37.5:1 crosslinker ratio), 2.5 mL of 1.5M of Tris (pH8.8), 4.7 mL of ddH_2O, 50µL of TEMED, 330 µL 0.05% APS.
18. Native –PAGE sample buffer 5X - the buffer was prepared by mixing 15.5 mL 1M Tris-HCl pH 6.8, 25 mL glycerol, 7 mL ddH_2O and a pinch of bromophenol blue.
19. Native PAGE running buffer (10x).- The buffer was prepared with 30g of Tris and 144g of glycine at 1L.
20. N-succinyl-AAPF stock solution – Prepared at a concentration of 20mM in DMSO and stored at −20°C.
21. PMSF (100mM) – The solution is prepared by adding 17.4 mg of PMSF per milliliter of isopropanol and stored at −20°C.
22. Sodium acetate, 3M – The buffer is prepared by dissolving 408g $C_2H_3NaO_2.3H_2O$ in 800mL ddH_2O. The pH is adjusted 4.8 or 5.2 with 3M acetic acid and volume to 1L.
23. TAE buffer (50x) - the buffer was prepared with 242 g/L of Tris, 5.71% (v/v) of glacial acetic acid, and 18.6 g/L of EDTA. 1x TAE buffer was subsequently prepared by diluting 20 mL of 50x TAE with 980mL of ddH_2O.
24. Wash buffers for His-Tag purification - The buffer contains the same components as *E. coli* lysis buffer, but added with various concentrations of imidazole.
25. SDS-PAGE gel-Stacking gel (4%) - the gel was prepared with 1 mL of 40% acrylamide/bis solution (37.5:1 crosslinker ratio), 2.52 mL of 0.5M of Tris (pH6.8), 6.36 mL of ddH_2O, 100µL of 10% SDS, 10µL of TEMED and 50uL of 10% (w/v) APS. Resolving gel (12%)- The gel was prepared with 3 mL of 40% acrylamide/bis solution(37.5:1 crosslinker ratio), 2.5 mL of 1.5M of Tris (pH8.8), 4.35mL of ddH_2O, 100 µL of 10% (w/v) SDS, 5µL of TEMED, and 50µL of 10 (w/v) APS.
26. SDS running buffer (5x) - the buffer was prepared with 15.1g/L if Tris, 72g/L of glycine and 5 g/L of SDS. No pH adjustment was necessary. 1x SDS running buffer was obtained by diluting 5x SDS running buffer to 1x.
27. Urea (5M) – The solution is prepared by dissolving 3 g urea in 5 mL of ddH_2O and adjust the volume to 10 mL.

28. Urea (8M) – The solution is prepared by dissolving 4.8g urea in 5 mL of ddH_2O and adjust the volume to 10 mL.
29. Sarkosyl (1% w/v) - dissolved 1g of sarkosyl in 100 mL of ddH_2O.
30. Tris-HCl (1.5M)- the solution is prepared by dissolving 182.1g Tris in 800ml ddH_2O. The pH can be adjusted to the desired pH using HCl. The final volume is 1L.
31. Fixing solution - the solution was prepared with 25% (v/v) isopropanol and 10% (v/v) acetic acid.
32. Coomasie staining solution - the solution was prepared with 60mg/L brilliant blue R in 10% (v/v) acetic acid. The solution was filtered to remove any undissolved precipitates.

Appendix B

List of reference ID for protein sequences used in the phylogenetic analysis in Chapter 2

Reference ID for subtilase protein sequences:

- gi|88856068|ref|ZP_01130729.1|
- tr|A5CTA5|A5CTA5_CLAM3
- tr|Q6ADS0|Q6ADS0_LEIXX
- gi|376002137|ref|ZP_09779984.1|
- gi|300863958|ref|ZP_07108872.1|
- gi|385810158|ref|YP_005846554.1|
- gi|218960895|ref|YP_001740670.1|
- sp|P00780|SUBT_BACLI
- sp|P29600|SUBS_BACLE
- sp|P28842|SUBT_BACS9
- sp|P7|P233147_EXPR_XANCP
- tr|B3WFP4|B3WFP4_9BASI
- gi|156233835|gb|ABU58618.1|
- sp|P06873|PRTK_TRIAL
- tr|Q8GB52|Q8GB52_9VIBR
- gi|88780865|gb|EAR12044.1|
- gi|428297570|ref|YP_007135876.1|

Reference ID for lipase GDSL protein sequences:

- gi|196195517|gb|EDX90476.1|
- gi|109699107|gb|ABG39027.1|
- gi|54303579|ref|YP_133572.1|
- gi|85820945|gb|EAQ42092.1|
- gi|126733610|ref|ZP_01749357.1|
- gi|161326464|gb|EDP97790.1|
- gi|350540616|gb|AAM97294.2|
- gi|133909726|emb|CAL99838.1|
- gi|53713020|ref|YP_099012.1|
- gi|338232230|gb|EGP07362.1|
- gi|342906297|gb|ABB37214.2|

- gi|91717830|gb|EAS84480.1|
- gi|118497486|ref|YP_898536.1|
- tr|Q9QUN4|Q9QUN4_RAT
- gi|114550048|gb|EAU52929.1|

Reference ID for GH13 protein sequences:

- gi|148273981|ref|YP_001223542.1|
- gi|88855957|ref|ZP_01130619.1|
- gi|72161237|ref|YP_288894.1|
- gi|126620231|gb|EAZ90952.1|
- gi|226375861|ref|YP_002789000.1|
- gi|50954846|ref|YP_062134.1|
- gi|145220475|ref|YP_001131184.1|
- gi|78486138|ref|YP_392063.1|
- gi|87284615|gb|EAQ76567.1|
- gi|158335445|ref|YP_001516617.1|
- gi|198257800|gb|EDY82108.1|
- gi|161897607|ref|YP_946704.2|
- gi|89890048|ref|ZP_01201559.1|
- gi|120406671|ref|YP_956500.1|
- gi|117927889|ref|YP_872440.1

Appendix C

Annotation of the HMMsearch results

Table C.1 Annotation of the HMMsearch result for subtilase

No	Sample ID	Sampling depth (m)	Sequence ID	KEGG description	NCBI-NR description	SWISSPROT description
1	232	5	167813433	Lxx17160 serine protease (*Leifsonia xyli*)	serine protease (*Leifsonia xyli* subsp. xyli str. CTCB07)	Thermophilic serine proteinase precursor - *Bacillus* sp. (strain AK1)
2	232	5	167774469	Tery_2365 peptidase S8 and S53, subtilisin, kexin, sedolisin (Trichodesmium erythraeum)	subtilisin-like protein *(Microcystis aeruginosa* PCC 7806)	Cell wall-associated protease precursor - *Bacillus subtilis*
3	232	5	167861678	Lxx17160 serine protease (*Leifsonia xyli*)	serine protease (marine actinobacterium PHSC20C1)	Thermophilic serine proteinase precursor - *Bacillus* sp. (strain AK1)
4	232	5	167890758	XCV0959 extracellular protease precursor *(Xanthomonas campestris pv. vesicatoria)*	extracellular protease precursor (*Xanthomonas campestris pv. vesicatoria* str. 85-10)	Extracellular protease precursor - *Xanthomonas campestris* pv. *campestris*
5	232	5	167865372	CMM_2260 putative serine peptidase, family S8 (*Clavibacter michiganensis* subsp. *michiganensis* K01362)	putative serine peptidase, family S8 (*Clavibacter michiganensis* subsp. michiganensis NCPPB 382)	Thermophilic serine proteinase precursor - *Bacillus* sp. (strain AK1)
6	232	5	167703364	SGR_2459 putative secreted subtilisin-like serine protease (*Streptomyces griseus* K01362)	serine proteinase (*Thermus* sp. Rt41A)	Extracellular serine proteinase precursor - *Thermus* sp. (strain Rt41A)
7	232	5	167718106	Rcas_2539 peptidase S8 and S53 subtilisin kexin sedolisin (*Roseiflexus castenholzii* DSM13941)	peptidase S8 and S53 subtilisin kexin sedolisin (*Roseiflexus castenholzii* DSM 13941)	Major intracellular serine protease precursor - *Bacillus subtilis*

Table C.1: Continue from previous page

No	Sample ID	Sampling depth (m)	Sequence ID	KEGG description	NCBI-NR description	SWISSPROT description
8	232	5	175873336	Lxx17160 serine protease (*Leifsonia xyli*)	serine protease (*Marine actinobacterium* PHSC20C1)	Thermophilic serine proteinase precursor - *Bacillus* sp. (strain AK1)
9	232	5	232_11123086 14602	Marky_2033 aqualysin 1 (EC:3.4.21.111) (*Marinithermus hydrothermalis)*	aqualysin 1 (*Marinithermus hydrothermalis* DSM 14884)	Extracellular serine proteinase *Thermus* sp. (strain Rt41A)
10	232	5	232_11132984 58479	blr3044 extracellular protease (EC:3.4.24.-); K01417 putative zinc metalloprotease (EC:3.4.24.-) (*Bradyrhizobium japonicum)*	hypothetical protein HMPREF9695_01075 (*Afipia broomeae* ATCC 49717)	Subtilisin BPN' (*Bacillus amyloliquefaciens*)
11	231	11.5	163430931	Lxx17160 serine protease (*Leifsonia xyli*)	serine protease (*Leifsonia xyli* subsp. xyli str. CTCB07)	Thermophilic serine proteinase precursor - *Bacillus* sp. (strain AK1)
12	231	11.5	163154518	Rcas_2539 peptidase S8 and S53 subtilisin kexin sedolisin	peptidase S8 and S53 subtilisin kexin sedolisin (*Roseiflexus castenholzii* DSM 13941)	Major intracellular serine protease precursor - *Bacillus subtilis*
13	231	11.5	163539195	Lxx17160 serine protease (*Leifsonia xyli*)	serine protease (*marine actinobacterium* PHSC20C1)	Thermophilic serine proteinase precursor - *Bacillus* sp. (strain AK1)
14	231	11.5	163120751	Lxx17160 serine protease (*Leifsonia xyli*)	serine protease (*marine actinobacterium* PHSC20C1)	Thermophilic serine proteinase precursor - *Bacillus* sp. (strain AK1)
15	231	11.5	163128715	XCV0959 extracellular protease precursor; K01362	extracellular protease precursor (*Xanthomonas campestris pv. vesicatoria* str. 85-10)	Extracellular protease precursor -*Xanthomonas campestris* pv. campestris
16	231	11.5	163497393	CMM_2260 putative serine peptidase, family S8; K01362 (*Clavibacter michiganensis* subsp. *Michiganensis)*	putative serine peptidase, family S8 (*Clavibacter michiganensis* subsp. *michiganensis* NCPPB 382)	Thermophilic serine proteinase precursor - *Bacillus* sp. (strain AK1)

Table C.1: Continue from previous page

No	Sample ID	Sampling depth (m)	Sequence ID	KEGG description	NCBI-NR description	SWISSPROT description
17	231	11.5	163343819	758118 LOC758118; similar to proprotein convertase subtilisin/kexin type 9 preproprotein (*Strongylocentrotus purpuratus*)	PREDICTED: similar to proprotein convertase subtilisin/kexin type 9 preproprotein (*Strongylocentrotus purpuratus*)	P12547|ORYZ_ASPOR Oryzin precursor - *Aspergillus oryzae*
18	231	11.5	168645997	BT_3889 subtilisin-like serine protease	subtilisin-like serine protease (*Polaribacter irgensii* 23-P)	P29599|SUBB_BACLE Subtilisin BL - *Bacillus lentus*
19	228	18	169958875	GK1517 subtilisin-type proteinase; K01362*(Geobacillus kaustophilus)*	putative serine peptidase precursor; putative fibronectin type III domain; putative signal peptide (*Candidatus* Cloacamonas acidaminovorans)	Thermophilic serine proteinase precursor - Bacillus sp. (strain AK1)

Table C.2: Annotation of HMMsearch result for Lipase 3

No	Sample ID	Sampling depth (m)	Sequence ID	KEGG description	NCBI description
1	232	5	167696458	Lipase family protein (*Tetrahymena thermophila* SB210)	Lipase family protein (*Tetrahymena thermophila* SB210)
1	231	11.5	163436800	Triacylglycerol lipase (*Monosiga brevicollis)*	Chain A, crystal structure of an extracellular triglyceride lipase (from a fungus *Rhizomucor miehei)*
2	231	11.5	163182232	hypothetical protein (*Monosiga brevicollis)*	predicted protein (*Monosiga brevicollis* MX1)

Table C.3: Annotation of the HMMsearch result for lipase GDSL

No	Sample ID	Sampling depth (m)	Sequence ID	KEGG description	NCBI-NR description	SWISSPROT description
1	232	5	167817518	lipolytic enzyme, GDSL family (*Saccharopolyspora erythraea*)	GDSL family (*Saccharopolyspora erythraea* NRRL 2338)	Glutamine-dependent NAD(+) synthetase (*Rhodopseudomonas capsulata*)
2	232	5	167780571	putative acyl-CoA thioesterase precursor; (*Desulfovibrio desulfuricans*)	putative acyl-CoA thioesterase precursor (*Desulfovibrio desulfuricans* G20)	Arylesterase (EC 3.1.1.2) (*Vibrio mimicus*)
3	232	5	167687840	GDSL-like lipolytic enzyme (*Francisella novicida*)	conserved hypothetical protein (*Francisella tularensis* subsp. *novicida* GA99-3548)	Esterase TesA (EC 3.1.1.1) (*Pseudomonas aeruginosa*)
4	232	5	167825930	lipolytic enzyme, GDSL (*Pseudoalteromonas atlantica*)	lipolytic enzyme, GDSL (*Pseudoalteromonas atlantica* T6c)	Galactolipase DONGLE, chloroplastic *Arabidopsis thaliana*
5	232	5	167882604	lipolytic enzyme, GDSL (*Pseudoalteromonas atlantica*)	GDSL-like lipase/acylhydrolase, putative (*Alcanivorax* sp. DG881)	UDP-N-acetylenolpyruvoylglucosamine reductase *Bradyrhizobium* sp. (strain ORS278)
6	232	5	167666764	lysophospholipase L1 and related esterase-like protein (*Pseudomonas mendocina*)	Lysophospholipase L1 and related esterase-like protein (*Pseudomonas mendocina*)	Lysophospholipase protein (*Herbaspirillum seropedicae*)
7	232	5	232_11132978 79091	arylesterase (EC:3.1.1.2)(*Candidatus Pelagibacter ubique*)	arylesterase (*Candidatus* Pelagibacter ubique HTCC1002)	Acyl-CoA thioesterase I (*Escherichia coli* (strain K12))
8	232	5	232_11132979 96613	lysophospholipase L1-like esterase (*Solitalea Canadensis*)	GDSL-like protein (*Prevotella stercorea* DSM 18206)	Platelet-activating factor acetylhydrolase IB(*Rattus norvegicus*)
9	232	5	232_11133161 02348	GDSL-like lipase/acylhydrolase domain protein (*Rubrivivax gelatinosus*)	GDSL-like protein (*Leptospira wolffii Serovar Khorat* Str. Khorat-H2)	Galactolipase DONGLE, chloroplastic (*Arabidopsis thaliana*)
10	232	5	232_11133160 24971	lipolytic protein GDSL (*Oligotropha carboxidovorans*)	lipolytic enzyme (*Bradyrhizobiaceae bacterium* SG-6C)	Esterase TesA (*Pseudomonas aeruginosa*)

Table C.3: *Continue from previous page*

No	Sample ID	Sampling depth (m)	Sequence ID	KEGG description	NCBI-NR description	SWISSPROT description
1	231	11.5	163414796	arylesterase (*Candidatus Pelagibacter ubique*)	arylesterase (*Candidatus* Pelagibacter ubique HTCC1002)	Arylesterase (EC 3.1.1.2)(*Vibrio mimicus*)
2	231	11.5	163429503	lipolytic enzyme, GDSL family (*Syntrophobacter fumaroxidans*)	GDSL family lipase (*Elusimicrobium minutum* lipolytic protein GDSL family	Acyl-CoA thioesterase I (EC 3.1.2.-) (Lecithinase B) (Lysophospholipase L1) (EC 3.1.1.5) (Protease I) (*Escherichia coli* (strain K12))
3	231	11.5	163235684	putative acyl-CoA thioesterase precursor; K01045 arylesterase (*Desulfovibrio desulfuricans*)	gi|78217865|gb|ABB37214.1| putative acyl-CoA thioesterase precursor (*Desulfovibrio desulfuricans G20*)	Arylesterase (EC 3.1.1.2) - (*Vibrio mimicus*)
4	231	11.5	163262734	hypothetical protein (*Thermobifida fusca*)	gi|84381226|gb|EAP97110.1| hypothetical protein JNB_16534 (*Janibacter* sp. HTCC2649)	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase (*Acyrthosiphon pisum* symbiotic bacterium)
5	231	11.5	231_1113297565557	GDSL lipolytic protein (*Pseudoalteromonas atlantica*)	gb|EPG65890.1| GDSL-like protein (*Leptospira wolffii Serovar Khorat* Str. Khorat-H2)	DNA mismatch repair protein MutS (*Synechococcus* sp. (strain CC9902))
6	231	11.5	231_1113297793571	lipolytic protein GDSL family (*Cellulophaga algicola*)	gb|ADV50112.1| lipolytic protein GDSL family (*Cellulophaga algicola* DSM 14237)	DNA mismatch repair protein MutS (*Synechococcus* sp. (strain CC9902))
7	231	11.5	231_1113289401511	GDSL-type lipase (*Desulfobacula toluolica*)	predicted lipase, GDSL-type (*Desulfobacula toluolica* Tol2)	Arylesterase (*Vibrio mimicus*)

Table C.4: Annotation of the HMMsearch results for GH13

No	Sample ID	Sampling depth (m)	Sequence ID	KEGG description	NCBI-NR description	SWISSPROT description
1	232	5	167733578	Putative α-glucosidase, glycosyl hydrolase family 13; (*Clavibacter michiganensis* subsp. *michiganensis* NCPPB 382)	Glycosyl hydrolase family 13 (*Clavibacter michiganensis* subsp. *michiganensis* NCPPB 382)	Maltase 2 precursor - *Drosophila virilis* (Fruit fly)
2	232	5	167822374	α-glucosidase. Glycosyl hydrolase family 13.; α-glucosidase (*Thermobifida fusca*)	α -glucosidase. Glycosyl Hydrolase family 13. (*Thermobifida fusca* YX)	Probable α-glucosidase - (*Sinorhizobium meliloti*)
3	232	5	167733580	α -glucosidase (*Clavibacter michiganensis* subsp. *michiganensis*)	α -amylase, catalytic subdomain (*Marine actinobacterium* PHSC20C1)	Maltase 2 precursor - *Drosophila virilis* (Fruit fly)
4	232	5	167687568	α -glucosidase. Glycosyl hydrolase family 13 (*Thermobifida fusca*)	α -glucosidase (*Azotobacter vinelandii* DJ)	Q45101|O16G_BACCO Oligo-1,6-glucosidase (*Bacillus coagulans)*
5	232	5	167667080	α -glucosidase (*Clavibacter michiganensis* subsp. *Michiganensis)*	α -amylase, catalytic subdomain (*marine actinobacterium* PHSC20C1)	Maltase 2 precursor (*Drosophila virilis*)(Fruit fly)
6	232	5	167715284	Trehalose synthase (*Chlorobium phaeovibrioides)*	Trehalose synthase (*Chlorobium phaeovibrioides* DSM 265)	Trehalose synthase (*Pimelobacter* sp. strain R48)
7	232	5	167853847	Conserved hypothetical lipoprotein (*Bacteroides fragilis* NCTC9343)	*A*-amylase (*Flavobacteria bacterium* BBFL7)	1,4-α-glucan-branching enzyme (*Vibrio vulnificus* (strain YJ016))
8	232	5	167817168	Glycogen branching enzyme; K00700 1,4-α-glucan branching enzyme (*Acidothermus cellulolyticus*)	1,4-α-glucan branching enzyme (*Acidothermus cellulolyticus* 11B)	1,4-α-glucan-branching enzyme (*Thermobifida fusca* (strain YX))

Table C.4: *Continue from previous page*

No	**Sample ID**	**Sampling depth (m)**	**Sequence ID**	**KEGG description**	**NCBI-NR description**	**SWISSPROT description**
9	232	5	167817172	Glycogen debranching enzyme GlgX; K02438 glycogen operon protein GlgX (*Mycobacterium vanbaalenii* PYR-1)	Glycogen debranching enzyme GlgX (*Mycobacterium vanbaalenii* PYR-1)	Glycogen operon protein glgX homolog (*Mycobacterium tuberculosis*)
10	232	5	175711712*	putative α-glucosidase (*Cyanothece* sp ATCC 51142)	α -amylase, catalytic domain subfamily, putative (*Microcoleus chthonoplastes* PCC 7420)	Probable α-glucosidase - (Sinorhizobium meliloti)
11	232	5	176222994*	α-amylase (*Acaryochloris marina* MBIC11017	α amylase (*Acaryochloris marina* MBIC11017)	α -amylase precursor (Geobacillus stearothermophilus)
12	232	5	175711710*	hypothetical protein (Cyanothece sp ATCC 51142)	α -amylase, catalytic domain subfamily, putative (*Microcoleus chthonoplastes* PCC 7420)	Probable α-glucosidase (*Sinorhizobium meliloti*)
13	232	5	175839179**	amylase, catalytic region (*Verrucomicrobiae bacterium* DG1235)	α -amylase, catalytic domain subfamily (*Verrucomicrobiae bacterium* DG1235)	α -amylase 2 (*Dictyoglomus thermophilum)*
14	231	11.5	163470518	Putative α-glucosidase, glycosyl hydrolase family 13 *(Clavibacter michiganensis* subsp*. Michiganensis)*	Putative α-glucosidase, glycosyl hydrolase family 13 (*Clavibacter michiganensis* subsp. *michiganensis* NCPPB 382)	Maltase 2 precursor (*Drosophila virilis*) (Fruit fly)
15	231	11.5	163360988	α -glucosidase. Glycosyl hydrolase family 13 (*Thermobifida fusca)*	α -glucosidase. Glycosyl Hydrolase family 13. (*Thermobifida fusca* YX)	Probable α-glucosidase (*Sinorhizobium meliloti*)
16	231	11.5	163470516	α -glucosidase *(Clavibacter michiganensis* subsp. *michiganensis*	α -amylase, catalytic subdomain (marine actinobacterium PHSC20C1)	Maltase 2 precursor (*Drosophila virilis* (Fruit fly))

* -sequence from 0.8 μm,**- sequence from 3.0 μm

Table C.4: *Continue from previous page*

No	Sample ID	Sampling depth (m)	Sequence ID	KEGG description	NCBI-NR description	SWISSPROT description
17	231	11.5	163427827	Glycosyl hydrolase family 13 α-glucosidase (*Thermobifida fusca*)	α -glucosidase (*Azotobacter vinelandii* DJ)	Oligo-1,6-glucosidase (*Bacillus coagulans)*
18	231	11.5	163488553	Trehalose synthase (*Leifsonia xyli* subsp. Xyli)	Trehalose synthase (*Leifsonia xyli* subsp. *xyli* str. CTCB07)	Trehalose synthase (*Pimelobacter* sp. (strain R48))
19	231	11.5	163239763	1,4-α-glucan branching enzyme (*Arthrobacter aurescens)*	1,4-α-glucan branching enzyme (*Arthrobacter aurescens* TC1)	1,4-α-glucan-branching enzyme (*Thermobifida fusca* (strain YX))
20	231	11.5	163277667	Glycogen operon protein GlgX (*Acidothermus cellulolyticus*)	Glycogen debranching enzyme GlgX (*Acidothermus cellulolyticus* 11B)	Glycogen operon protein glgX homolog (*Mycobacterium tuberculosis)*
21	231	11.5	163239763	Glycogen operon protein GlgX (*Mycobacterium vanbaalenii*)	Glycogen debranching enzyme GlgX (*Mycobacterium vanbaalenii* PYR-1)	Glycogen operon protein glgX homolog (*Mycobacterium tuberculosis)*
22	231	11.5	167541905*	α-amylase, catalytic region *(Thiomicrospira crunogena)*	α-amylase, amylosucrase (*Synechococcus* sp. WH 5701)	Amylosucrase (*Neisseria polysaccharea*
23	231	11.5	167562621*	α-glucosidase *(Cyanothece* sp ATCC 51142)	α -amylase, catalytic domain subfamily, putative (*Microcoleus chthonoplastes* PCC 7420)	Probable α-glucosidase (*Sinorhizobium meliloti*)
24	231	11.5	167588808*	amylase (*Acaryochloris marina)*	α-amylase (*Acaryochloris marina* MBIC11017)	α -amylase precursor (*Geobacillus stearothermophilus*)
25	231	11.5	168627242**	α-amylase, catalytic region *(Thiomicrospira crunogena)*	α-amylase, amylosucrase (*Synechococcus* sp. WH 5701)	Amylosucrase (*Neisseria polysaccharea)*
26	228	18	178055561	trehalose synthase (*Chlorobium phaeovibrioides* DSM 265)	trehalose synthase (*Chlorobium phaeovibrioides* DSM 265)	Trehalose synthase (Pimelobacter sp. (strain R48))
27	228	18	167494038	trehalose synthase (*Chlorobium phaeovibrioides* DSM 265)	trehalose synthase (*Chlorobium phaeovibrioides* DSM 265)	Trehalose synthase (*Pimelobacter* sp. (strain R48))

Appendix D

Table D.1: Final concentration of ammonium sulfate: Percentage saturation at 0°C*

Initial concentration of ammonium sulfate (percentage saturation at 0 °C)	Percentage saturation at 0 °C																
	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
	Solid ammonium sulfate (g) to be added to 1 l of solution																
0	106	134	164	194	226	258	291	326	361	398	436	476	516	559	603	650	697
5	79	108	137	166	197	229	262	296	331	368	405	444	484	526	570	615	662
10	53	81	109	139	169	200	233	266	301	337	374	412	452	493	536	581	627
15	26	54	82	111	141	172	204	237	271	306	343	381	420	460	503	547	592
20	0	27	55	83	113	143	175	207	241	276	312	349	387	427	469	512	557
25		0	27	56	84	115	146	179	211	245	280	317	355	395	436	478	522
30			0	28	56	86	117	148	181	214	249	285	323	362	402	445	488
35				0	28	57	87	118	151	184	218	254	291	329	369	410	453
40					0	29	58	89	120	153	187	222	258	296	335	376	418
45						0	29	59	90	123	156	190	226	263	302	342	383
50							0	30	60	92	125	159	194	230	268	308	348
55								0	30	61	93	127	161	197	235	273	313
60									0	31	62	95	129	164	201	239	279
65										0	31	63	97	132	168	205	244
70											0	32	65	99	134	171	209
75												0	32	66	101	137	174
80													0	33	67	103	139
85														0	34	68	105
90															0	34	70
95																0	35
100																	0

* Adapted from (Burgess, 2009a)

Appendix E

Multiple sequence alignment of Subt9195, Subt5372, Subt8715 and Subt4518 with matches sequence from NCBI and Swiss-Prot databases

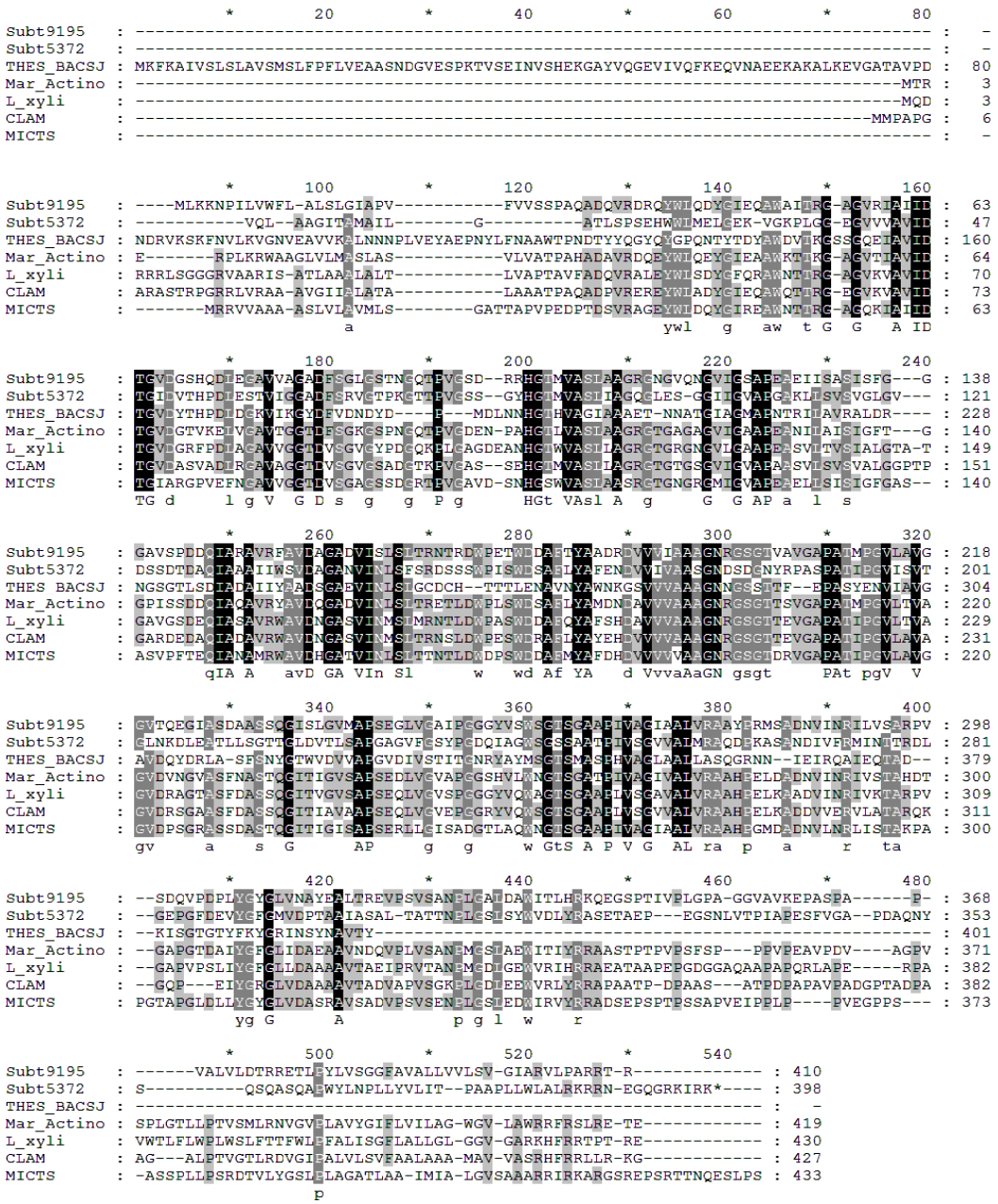


Figure E.1: Multiple alignment of the deduced amino acid sequences of Subt9195 and Subt5372 with matches from NCBI and Swiss-Prot database. Reference sequence uses for alignment are : Mar_actino; (gi|88856068|) Serine protease marine actinobacterium PHSC20C1, L_.xyli; (gi|50951775) serine protease *Leifsonia xyli* CTCB07, CLAM; (gi|148273444|) Putative serine peptidase family S8 *Clavibacter michiganensis* (strain NCPPB 382), MICTS; (gi323359447); Subtilisin-like serine protease *Microbacterium testaceum* (strain StLB037) and THES_BACSJ; (Q45670) Thermophilic serine proteinase *Bacillus* sp. (strain AK1).

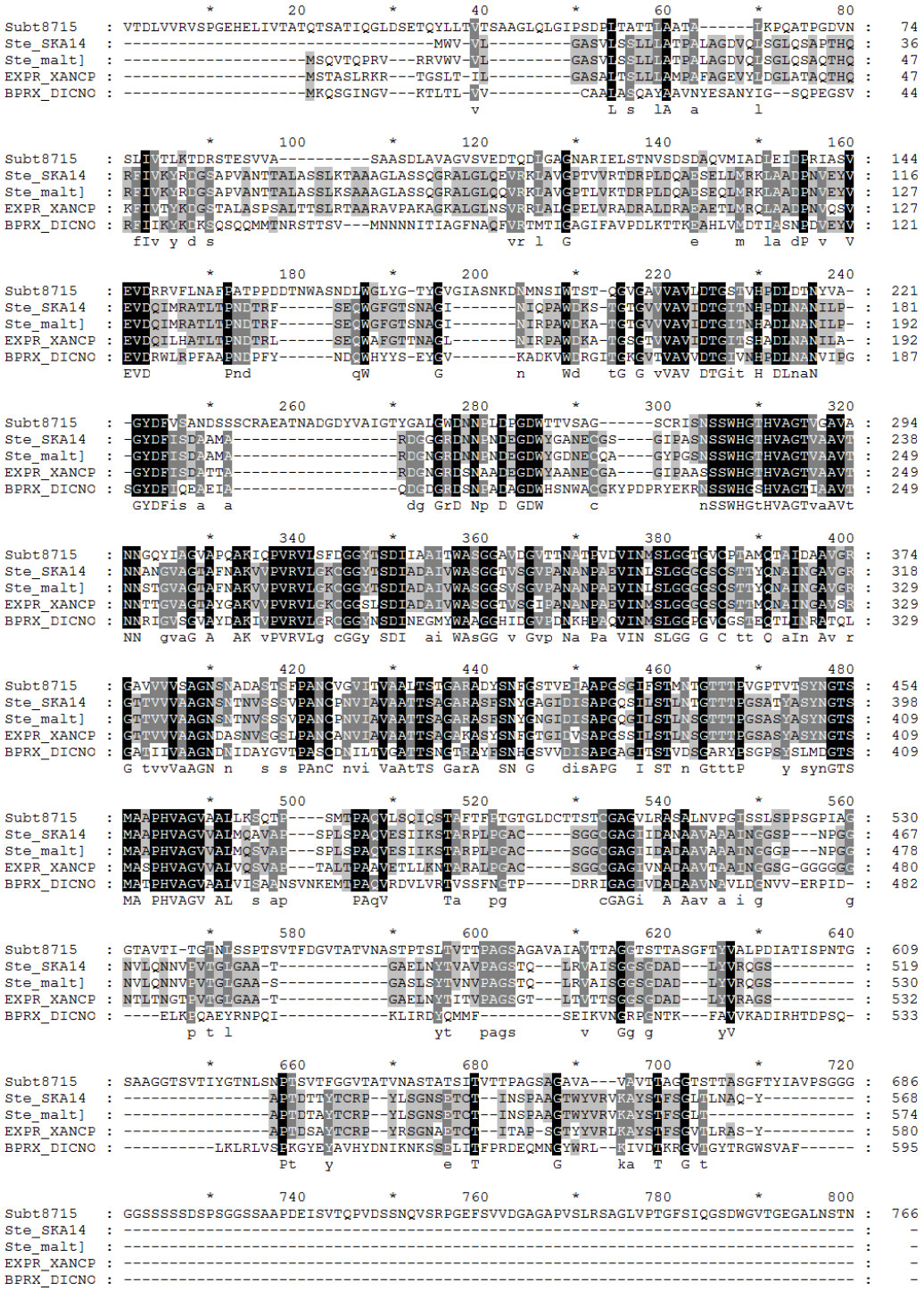


Figure E.2: Multiple sequence alignments of the deduced amino acid sequences of Subt8715 (766aa) and matched sequences from NCBI and Swiss-Prot databases. Reference sequences used in the alignment are: Ste_SKA14 (ZP_05135285) extracellular protease *Stenotrophomonas* sp. SKA14, Ste_maltj: (CBI67289) minor extracellular protease *Stenotrophomonas maltophilia*, EXPR_XANCP: (P23314) Extracellular protease Xanthomonas campestris pv. campestris (strain ATCC 33913, BPRX_DICNO: (P42780) Extracellular subtilisin-like protease *Dichelobacter nodosus*.

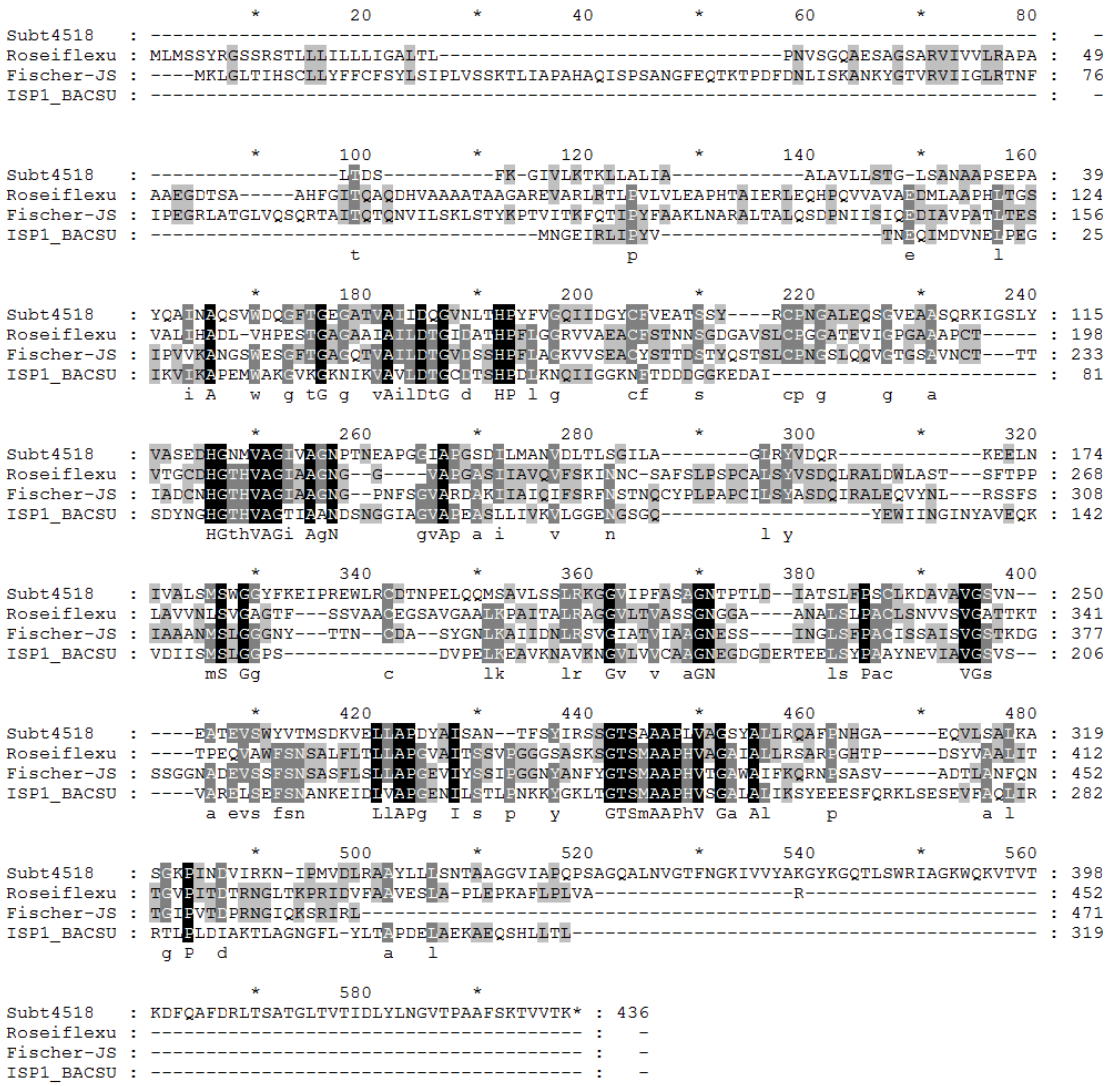


Figure E.3: Multiple sequence alignments of the deduced amino acid sequences of Subt4518 and matched sequences from NCBI-nr and Swiss-Prot databases. Reference sequences used in the alignment are: Roseiflexu (YP_001432636) peptidase S8/S53 subtilisin kexin sedolisin *Roseiflexus castenholzii* DSM 13941, Fischerella-JS: (ZP_08988107) peptidase S8 and S53 subtilisin kexin sedolisin *Fischerella* sp. JSC-11, ISP1_BACSU: (P11018) Major intracellular serine protease *Bacillus subtilis* (strain 168)

Appendix F

Table F.1: The spreadsheet for identification of the targeted clone from the sequence-based screening in the metagenomic clone library

HOTEL_BARCODE	HOTEL_COLUMN_BARCODE	CONTAINER_BARCODE	GROWTH_BARCODE	LIBRARY_NAME	PROJECT_NAME
HOT000025MS	HOT500025MS	DD0003FJLNU	BB0222752AB	ANTRC227-G-01-4-6KB	Antarctica 2007
HOT000025Q1	HOT300025Q1	DD0003FN3NU	BB0229985AB	ANTRC228-G-01-3-4KB	Antarctica 2007
HOT000025MC	HOT200025MC	DD0003FNMNU	BB0230849AB	ANTRC228-G-01-3-4KB	Antarctica 2007
HOT000025DV	HOT100025DV	DD0003FZZNU	BB0227632AB	ANTRC229-G-01-3-4KB	Antarctica 2007
HOT000025DV	HOT100025DV	DD0003G00NU	BB0227633AB	ANTRC228-G-01-3-4KB	Antarctica 2007
HOT000025MS	HOT400025MS	DD0003HFUNU	BB0232543AB	ANTRC235-G-01-4-6KB	Antarctica 2007
HOT000025MC	HOT400025MC	DD0003HP4NU	BB0235893AB	ANTRC230-G-01-3-4KB	Antarctica 2007
HOT000025MC	HOT400025MC	DD0003HPCNU	BB0235895AB	ANTRC230-G-01-3-4KB	Antarctica 2007
HOT000025MC	HOT400025MC	DD0003HPDNU	BB0235873AB	ANTRC230-G-01-3-4KB	Antarctica 2007
HOT000025MC	HOT500025MC	DD0003HPVNU	BB0236508AB	ANTRC230-G-01-3-4KB	Antarctica 2007
HOT000025MC	HOT500025MC	DD0003HPWNU	BB0236810AB	ANTRC230-G-01-3-4KB	Antarctica 2007
HOT000025MC	HOT500025MC	DD0003HPXNU	BB0236530AB	ANTRC230-G-01-3-4KB	Antarctica 2007
HOT000025MC	HOT400025MC	DD0003HQGNU	BB0236529AB	ANTRC230-G-01-3-4KB	Antarctica 2007
HOT000025MC	HOT300025MC	DD0003HQHNU	BB0236507AB	ANTRC230-G-01-3-4KB	Antarctica 2007
HOT0000262X	HOT4000262X	DD0003HUONU	BB0235510AB	ANTRC232-G-01-4-6KB	Antarctica 2007
HOT0000262X	HOT3000262X	DD0003HUPNU	BB0232377AB	ANTRC232-G-01-4-6KB	Antarctica 2007