MOLECULAR STUDIES AND BIOINFORMATIC ANALYSIS OF INDUSTRIALLY IMPORTANT PROTEASE CODING GENES IN METAGENOMIC EXTRACTS AND PURE ISOLATES FROM EXTREME ENVIRONMENT

THESIS
Submitted To

THE FACULTY OF LIFE & ENVIROMENTAL SCIENCES
UNIVERSITY OF PESHAWAR

In fulfillment for the degree

of

DOCTOR OF PHILOSOPHY

by

NOREEN AZHAR

CENTER OF BIOTECHNOLOGY & MICROBIOLOGY
UNIVERSITY OF PESHAWAR
2010

DEDICATION

This thesis is dedicated to My Family members,

especially to My Parents , My Husband and

My kids

for their Prayers, Unprecedented Support

and Eternal love.

ACKNOWLEDGEMENT

*All praises to Almighty **Allah**, the creator, the dominant, the most supreme, the self-existing and all sustaining, whose grace and mercy enabled me to accomplish this piece of work by sacredness of Holy Prophet **Muhammad (PBUH)** who is forever a torch of guidance and acknowledge for humanity as a whole.*

*Expression some times does not reflect the feelings operative behind it, and feelings most of the time does not find words proper for their manifestation. With deep sense of acknowledgement, I would like to express my humble gratitude to my worthy supervisor **Prof. Dr. Ghosia Lutfullah**, Centre of Biotechnology & Microbiology, University of Peshawar & co-supervisor **Dr. Muhammad Afzal Ghauri** (Principal Scientist), Head Industrial Biotechnology Division, National Institute for Biotechnology and Genetic Engineering (NIBGE) Faisalabad for their learned guidance, inestimable counsel, captivating attitude in the procedural matters as well as for their marvelous supervision in the concerned complication.*

*I am extremely indebted to **Prof. Dr. Bashir Ahmed**, Director, Centre of Biotechnology & Microbiology, University of Peshawar, for his great help, encouragement also for providing me a chance to work at NIBGE laboratories for research and further extension in the working period.*

*I would like to express my deep appreciation, regards and thanks to **ICGEB** (International Center for Genetic engineering and biotechnology ,Italy) for awarding me PhD-sandwich research fellowship and making it possible for me to carryout research activities at NIBGE, Faisalabad. It is an honour to acknowledge the kind and affectionate role of **Dr. Zafar M. Khalid** (Director, NIBGE Faisalabad) for permitting me to work as collaborative student under the supervision of **Dr. Muhammad Afzal Ghauri** and providing me all the required facilities and appreciating me during the period of my stay at NIBGE,Faisalabad .*

*I am glad to mention here and say thanks to all those who made it possible for me. I would pay special regards to **Miss. Nasrin Akhtar** (JS) for always being there to support, help, guidance and technically assist when ever I needed. My heartfelt regards are due to **Mrs. Shazia khaliq** (JS) and **Dr. Kalsoom Akhtar** (PS) for their productive discussions and valuable suggestions throughout my research work at NIBGE. I would also like to extend sincere gratitude to all my teachers and staff members of this Institute.*

*I feel great pleasure to express my appreciation for all my division fellows at NIBGE Faisalabad, **Ms. Ijaz Bano, Ms. Fatima, Ms. Nabeela, Mr.Rizwan Haider, Mr. Ilyas Khan**, **Mr**. **Farooq** and **Mr. Abdud Tawab** who helped , encouraged and supported me in every way to achieve this goal. I would never forget to acknowledge **Mr.Faqir Muhammad**, **Mr. Habib-ur-Rehman, Mr. Amir, Mr. Niaz** and **Mr. Ejaz**, without their assistance this task would have never been easier.*

*I feel immense pleasure to express my deepest gratitude and sincere thanks to all my research fellows, **Mrs. Farhat Humayun**, **Ms. Seema Gul, Mr. Zahid Khan** and **Mr. Roshan Ali** for their contribution to the success of this study.*

More importantly I want to pay special tribute to my great parents, who have consistently encouraged & motivated me for my every educational accomplishment, for endless spiritual and unconditional support they gave me all those years to complete my research work ambitiously. Also, the eternal love, care and moral support of my brothers, ***Mr. Taqweem Azhar****,* ***Mr.Fahim Azhar*** *and* ***Mr.Wasim Azhar*** *helped me to continue and complete my thesis.*

Additionally this dissertation would have never been possible without my husband ***Dr. Muhammad Asim*** *who encouraged and supported me to carry out the research activities at NIBGE, Faisalabad. I am extremely greatful to him for his patience and every possible help during the course of my studies.*

Last but not the least, my whole hearted love for my very loving kids ***Muhammad Umar Asim*** *and* ***Muhammad Ayyan Asim*** *who sacrificed a lot during my studies.*

Finally, I pray Almighty Allah to enable me to serve humanity in the best possible manner.

Noreen Azhar.

CONTENTS

LIST OF FIGURES

Page No.

LIST OF TABLES

LIST OF ABBREVATIONS

PCR	**Polymerase chain reaction**
DNA	**Deoxyribonucleic acid**
rDNA	**Ribosomal Deoxyribonucleic acid**
RNA	**ribose nucleic acid**
EC	**Enzyme classification**
kDa	**Kilo dalton**
ORF	**Open Reading Frame**
BPT-05	**Isolate code for *Bacillus amyloliquefaciens***
BPT-06	**Isolate code for *Bacillus cereus***
BPT-08	**Isolate code for *Staphylococcocus arlettae***
BPT-20	**Isolate code for *Bacillus licheniformis* KL-176**
BPT-23	**Isolate code for *Bacillus licheniformis* GXN151**
BPT-25	**Isolate code for *Bacillus pumilus***
AP	**Alkaliphilic**
HP	**Halophilic**
W/V	**Weight by volume**
Sa	**Soil sample a**
Sb	**Soil sample b**
Sc	**Soil sample c**
EDTA	**EthylinediamineTetraacetic acid**
TE Buffer	**Tris EDTA**
3D	**Three Dimentional**
PDB	**Protein Data Bank**
BLAST	**Basic Local Alignment Search Tool**

FASTA	**Fast Alignment**
PG05	**Protease gene 05**
PG08	**Protease gene 08**
PG20	**Protease gene 20**
PG23	**Protease gene 23**
PG25	**Protease gene 25**
PGSc	**Protease gene from soil sample c**

Amino Acid	Three-Letter-Code	One- Letter-Code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

List of Publications

1. Ghosia Lutfullah, **Noreen Azhar,** Farhat Amin, Zahid Khan, M. Kamran Azim, Khalida Shouqat, Sajid Noor, Mohd Rizwan. Structural Bioinformatics of *Vibrio cholerae* Aminopeptidase A. Protein & Peptide Letters, 2009, Vol. 16, No. 1, 36-45.
2. Ghosia Lutfullah, Farhat Amin, Zahid Khan, **Noreen Azhar**, M. Kamran Azim, Sajid Noor, Khalida Shoukat. Homology Modeling of Hemagglutinin/Protease [HA/P (vibriolysin)] from Vibrio Cholerae: Sequence Comparision, Residue Interactions and Molecular Mechanism. Protein J, 2008, 27:105–114.
3. **Noreen Azhar**, M. Afzal Ghauri, Nasrin Akhtar, Ghosia Lutfullah. Microbiological and 16S rRNA analysis for identification of five bacterial isolates from pesticide factory waste. Accepted in Pakistan Journal of Botany. Ref No.1120/CPB.
4. **Noreen Azhar**, Nasrin Akhtar, M. Afzal Ghauri, Ghosia Lutfullah. Molecular Cloning, sequencing and bioinformatics of Halophilic alkaline Serine Protease Gene from *B. licheniformis* strain GXN 151. Submitted in Canadian Journal of Microbiology.
5. **Noreen Azhar**, Nasrin Akhtar, M. Afzal Ghauri, Ghosia Lutfullah. Screening, Isolation, molecular characterization of protease gene from metagenomic soil samples.Submitted in Biotechnology Letters.
6. **Noreen Azhar,** Nasrin Akhtar, M. Afzal Ghauri, Ghosia Lutfullah. Biodiversity and phylogenetic analysis of culturable bacteria from pesticide factory waste and their importance. To be Published

SUMMARY

Proteases are large group of ubiquitous enzymes having small size and are compact structures. They are very important industrial enzymes possessing wide range of applications in the detergent, food, leather and pharmaceutical industries. Microbial proteases are known for their key role in biotechnological applications. Metagenomic techniques set a new dimension in exploring novel genes. To maximize industrial enzyme production, the host organism is generally manipulated to carry multiple copies of protease gene. This can be accomplished by cloning the gene on a replicating plasmid, but preferentially this is achieved by amplification of the gene on the chromosome, as such amplification offers the more stable alternative.

Current studies deal with the study of biodiversity contained in indigenous soil of varied nature for the exploration of industrially important protease gene. Moreover, protease genes have been PCR amplified, cloned and sequenced from various cultured microorganisms and metagenomic soil samples, molecular aspects of these genes were studied using different molecular biology techniques, prior to this they were tested for protease activity on casein agar plates. Various bioinformatics tools have been employed for the characterization of these genes. Also protein structures prediction was performed to know the structural and functional features of sequenced protease gene.

The indigenous isolates were used in this study, which were obtained from NIBGE culture collection. Time course studies revealed that current studies showed that *Bacillus amyloliquefaciens* (BPT- 05), *Bacillus cereus* (BPT-06), *Staphylococcus arlettae* (BPT-08), *Bacillus licheniformis* strain KL 176(BPT-20) and *Bacillus pumilus* (BPT-25) were moderate halophiles, while *Bacillus licheniformis* strain GXN 151 (BPT-23) presented extreme halophlic characteristics. BPT05 was of

extreme alkaliphilic nature while all other isolates were alkalotolerent , when tested in halophilic and alkaliphilic media for growth. Genomic DNA were extracted and amplified by PCR using six pairs of protease gene specific primers from the BPT-05, BPT-06, BPT-08, BPT-20, BPT-23 and BPT-25 and metagenomic soil sample. Fragment of 1101bp (BPT-05), 1156bp (BPT-08), 819bp (BPT-20), 950bp (BPT-23), 810bp (BPT-25), 429bp (S_C) were sequenced.

After sequencing, BLAST (NCBI) search was performed to determine nucleotide matches to the world databases. Pair wise sequence comparison of sequenced protease genes showed that *B.amyloliquefaciens* PG05 has 94% homology with *B.licheniformis* keratinolytic protease gene, *S.arlettae* PG08 has 93% homology with *B.licheniformis* protease gene, *B.licheniformis* strain KL176 PG20 and *B.pumilus* PG25 has 99% and 98% homology respectively with *B.pumilus* organic solvent tolorent protease gene, while *B.licheniformis* strain GNX 151 PG23 has 68% homology with *B.pumilus* alkaline serine protease gene.

ORF and Cognitor results showed that +2 and +2 open reading frame in PG05 has significant homology with subtilisin like serine protease gene, ORF -1 matches with Molybdopterin biosynthesis enzyme while other ORFs did not give any matches, indicating that these ORFs are novel. PG-08 has ORF +2 , PG20 and PG23 has ORF+1 ,PG 25 has ORF +3 matching with subtilisin like serine protease gene, while the ORFs which did not had any matches are assumed to be novel. Sequenced protease genes were then translated into respective proteins. Sequence homology searches for all translated protein sequences were carried out by using protein BLAST at NCBI. PG05 and PG08 showed 65% and 66% homology to subtilisin calsberg from *B. Subtilis* respectively. PG20 was 98% homologous to Keratinase precursor protein from *B.pumilus.* PG23 had 43% homology to serine alkaline

protease preproprotein from *B.pumilus.* PG25 showed 98% homology to organic solvent tolerant protease of *B.pumilus.* where as PGSc was 82% homologus to subtilisin protein from *B.pumilus*. There is found His active site motif (HGTHVAGTIAA) in PG05, PG08, PG20 and PG25. There is considerable active site conservation in active site region of these proteases. This motif is not present in PG23 and PGSc. Protein phylogenetic results show that PG05, PG08, PG20, PG23, PG25 and PGSc have common ancestary. According to the results PG05 &PG08, PG20 & PG25, PG23 & PGSc lie close to each other having common ancestral genes with little divergence. On the basis of sequence similarity comparative protein models were built.The tertiary structure of PG05 comprises of α / β/ α structural topology ,PG08 , PG20 and PG 25 has 2 β α β units while PG23 and PGS_C has 1 β α β unit. Ramachandran plot indicates that 88.4%, 87.8%, 85.4%, 83.6%, 84.7%, 91.9% of the main-chain dihedral angles are found in the most favored regions for PG05, PG08, PG20, PG23, PG25 and PGSc respectively. There are zero residues in disallowed region for all protein models except for PG23 which has 0.8% residues in the disallowed region. Prosa results shows z-score values below zero for all protein models. This is the indication of good quality model. All these nucleotide and protein structure analysis of proteases will help us better understand their catalytic activity, substrate binding and structural stability, this can help us in enhancement of their industrial and biotechnological application.

CHAPTER

1.0

INTRODUCTION

1.1. Microorganisms of industrial significance

Microbes have the ability to undergo diverse chemical transformations by the process of metabolism. Microbes could be utilized in exploring new industrial products. Microbes have the ability to convert cheap natural wastes into profitable daily use products. There are different ways by which microbes can be exploited industrially, this includes:

a. production of different products by fermentation
b. conversion of simpler compounds to complex once
c. Formation of various industrial harvests like enzymes and antibiotic etc.
d. Microbes as foodstuff.

Microbes including different bacteria, viruses, fungi, algae etc produce chemical substances that help in devising and assembling proteins with having unique properties that are of significant industrial relevance.

A much generalized equation depicting any industrial process can be presented as under:

Microorganism + Substrate $\longrightarrow$ Product

According to this equation microorganisms are one of the key components, which could be in the form of whole cell, endogene compound, a gene or a group of genes, an enzyme or a group of enzymes etc. In principle, any one of above given entity may act as a “Biocatalyst” in any industrial process. Each one of the above given components must possess some basic characteristics like, microorganisms of the reaction should be able to produce enough amount of product, it should be fast growing and non pathogenic. The substrates should be rich enough to support growth of microorganisms and are readily available. In case of product, there should be efficient recovery methods. Moreover the process could be up scaled to a pilot level or further for commercial exploitation.

1.2.Microbial biodiversity

Microbial biodiversity defines the diversity of all life forms including biotic communities and biotic processes that acknowledge the order at genetic, taxonomic and ecosystem levels. molecular and chemical diversity of these microbes is of + in nature.

1.3.Methods used for studying microbial biodiversity

While investigating microbial diversity within an environment two types of studies can be conducted, one is culture dependent and another is culture independent (Juck *et al*., 2000). Bacterial isolation from environmental samples using culture medium then leads to nucleic acid extraction from pure cultures involves culture dependent techniques. One important shortcoming in exploring microbial diversity is the concern of viable but non-cultivable organisms. Phenotypic characterization of isolated strains is used to determine diversities in bacterial communities. But this is only applicable on cultured bacteria. Culturing of majority of bacterial species is still not achievable even with advance microbiological culture techniques. Characterization of microbial strains using conventional methods is arguable, which depends upon particular environmental conditions for the growth of strains.(Bakonyi *et al*., 2003). These methods will result in aberrant changes in microbial population, when phenotypic diversity is taken into consideration most bacteria will be excluded only a small number of isolated bacteria will represent the total bacterial biodiversity. It can be assumed that uncultured species may possess important role in life like known bacteria.DNA is directly isolated from environmental samples in culture independent methods, direct DNA extraction accounts for a large proportion of uncultured microbes. Isolated DNA or cDNA from extracted

RNA is then amplified by PCR. Or else, cloning and sequencing of the amplified products is carried out to specify bacterial species present in the sample.

1.4. Enzymes

Living cell is the hot spot of remarkable biochemical activities commonly known as metabolism. This results in important chemical and physical changes in the living organisms. Including development and replacement of tissues, transfer and production of energy from food , excretion of waste etc.

Enzymes are important form of catalytic proteins in man, animals, plants and all life forms for their life and growth. Enzymes are the biocatalysts that enhance the speed of a biochemical reaction. They are of great value over traditional chemical catalysts, as they are clean, cheaper having easy recycling nature. However, to enable enzymes to be used in industrial processes, exploration of enzymes is needed to determine that whether they have the potential to execute a specific reaction under a set of industrial reaction condition (De-Kreij et al., 2000).

Enzymes are potentially much useful because of their enormous catalytic activities and high degree of specificity. They their utmost applications in various industrial sectors, which includes detergent, food, pharmaceutics, diagnostics, and fine chemical industries. More than 3000 different enzymes are described to date and the majority have been isolated from mesophilic organisms. Most of the industrially important enzymes are those that catalyze the digestion or hydrolysis of large molecules like starch, cellulose and proteins. The enzymatic action results in the production of simpler substances from these complex molecules. Of the industrial enzymes (with scale of 1 billion dollars

widely), 75% are hydrolytic in nature (Kumar and Takagi., 1999). These enzymes are specific in their action. They function only in specific conditions of temperature and pH. Hunt for new microbial resources are frequently in practice keeping in view biodiversity of microbes. Extremophilic microbes offer rich source of enzymes having unique properties which can be employed in novel processes (Kumar and Takagi, 1999) . Exploration of new microbial source is continually in process taking biodiversity into consideration.

1.4.1. Distribution of industrial enzymes

The pie chart below shows the contribution of different industrial enzymes showing the maximum contribution of proteases in the world market i.e. 59%

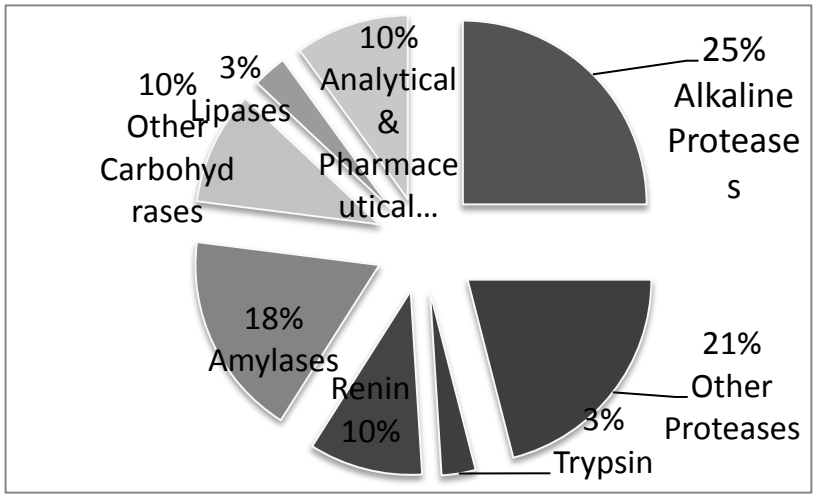


Fig.1.1. The share of different industrial enzymes to worldwide sale; the dark portion showes the total percentage of protease sale (Rao et al ., 1998).

1.5.Proteases

Proteases are the group of hydrolytic enzymes degrading proteins to peptides or free amino acids (Chen et al., 1991). Therefore they are put into hydrolases class and categorized in the subclass peptide hydrolases or peptidases (Ellaiah *et al*., 2002).

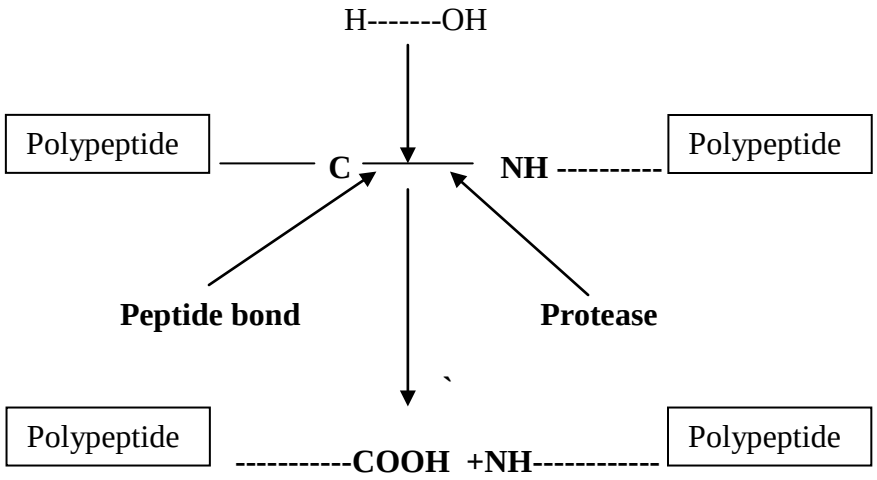


Fig.1.2. Protease catalysis of peptide bonds (proteolysis) (Chen et al., 1991).

Proteases are large group of ubiquitous enzymes having small size and are compact structures (Polgar, 1989). These enzymes have diverse applications in biological processes including metabolism , pathogenesis, protein modification and expression of genes etc (Rao et al., 1998). Their characterization is complex because of their diversity of action and structure. (Beynon et al., 1989; Rao et al ., 1998. Proteases have larger share than any other enzyme in global market, making them most important group of industrial enzymes (layman, 1986). These enzymes find a wide range of applications in the detergent, food, leather and pharmaceutical industries etc , they were brought in as

detergent additives in 1914 (Yang et al., 2000;Gupta et al., 2002b). So, proteases are of utmost significance as commercial enzymes (Bhosale et al., 1995).

1.5.1. Classification of proteases

Proteases are classified as hydrolases (Rao et al., 1998). Grouping of proteases is carried out on the basis of following properties;

a. On the basis of origin as animal, plant and microbial proteases.

b. As exo and endoenzymes on the basis of cleavage site

c. As serine, aspartic, cysteine and metallo proteases with regard to chemical nature of active site and catalytic mechanism.

d. On the basis of structural evolutionary relationship they are divided into clans and families.

Table 1.1. Classification of proteases (Rao et al., 1998)

Protease	Ec no.
1. Exopeptidase	3.4.11
1.1. Aminopeptidase	3.4.14
- Dipeptidyl peptidase	3.4.14
- Tripeptidyl peptidase	3.4.16-
1.2. Carboxypeptidase	3.4.18
- Serine type protease	3.4.16
- Metalloprotease	3.4.17
- Cysteine type protease	3.4.18
- Peptidyl dipeptidase	3.4.15
1.3. Omega peptidase	3.4.13
2. Endopeptidase	3.4.19
- Serine protease	3.4.19
- Cysteine protease	3.4.21-
- Aspartic protease	3.4.34
- Metallo protease	3.4.21
- Endopeptidase of unknown catalytic mechanism	3.4.22
	3.4.24
	3.4.99

*EC stands for enzyme classification.

1.5.2.1. Exo-peptidases

Exoproteases cleaves the bonds on C and N terminal of a polypeptide chain, they are grouped as aminopeptidases and carboxypeptidases with respect to their site of action (Rao et al., 1998).

i. Aminopeptidases

Peptidases that act on free amino terminal (N-terminal) of a protein chain releasing a mono, di or tripeptide. Methionine is an aminoacid normally present at the N-terminal of hetrologously expressed proteins is removed by aminopeptidases.

ii. Carboxypeptidases

As the name indicates the carboxypeptidases act at C terminals of the polypeptide chain and liberating a single amino acid or a dipeptide. (Rao et al., 1998).

1.5.2.2. Endopeptidases

Endopeptidases are characterized by their preferential action at the peptide bonds in the inner regions of the polypeptide chain away from the N and C termini. The presence of the free amino or carboxyl group has a negative influence on enzyme activity. The endopeptidases are divided into four subgroups based on their catalytic mechanism, (i) serine proteases, (ii) aspartic proteases, (iii) cysteine and (iv) metalloproteases (Rao et al., 1998).

i. Serine proteases

Serine group is present at the active site which is the distinguishing character of serine proteases and covalently binded with substrates and inhibitors. They are ubiquitous and extensively found in viruses, bacteria, and eukaryotes, signifying their critical

importance to the organisms. Their optimal activity is shown at neutral and alkaline pH, in the range between pH 7 and 11, isoelectric point is around pH 9 with molecular weights ranging from 15 to 30 kDa. Subtilisins are the best produced serine alkaline protease by *Bacillus* spp. representing the second largest family of serine proteases. They are among the extensively studied and best implicit enzymes (Rao et al., 1998).

ii. Aspartic proteases

Aspartic acid proteases are acidic endopeptidases that entail on aspartic acid residues for their catalytic activity showing their maximal activity at low pH with isoelectric point between pH 3-4.5 and molecular weight ranging from 30-45 kDa. There are three families of acidic proteases i.,e. pepsin, retropepsin and enzymes from pararetroviruses. Pepstatin is the inhibitor of aspartic acid proteases (Rao et al., 1998).

iii. Cysteine / thiol proteaeses

Both prokaryotic and eukaryotic organisms have cysteine proteases having cysteine or HCN at their active site, which play a critical role in their activity. Cysteine or and can react with a variety of reagents; heavy metals, iodoacetate, N-ethyl-maleimide etc (Kenny, 1999). Based on their side-chain specificity, they are broadly divided into four groups: (i) papain-like, (ii) trypsin-like, (iii) specific to glutamic acid and (iv) others. Papain is the best-known cysteine proteases. Cysteine proteases have neutral pH optima (Rao et al., 1998).

iv. Metalloproteases

Metallo protease requires a divalent metal ion for their activity.

Based on the specificity of their action, metalloproteases can be divided into 4 groups, (i) neutral, (ii) alkaline, (iii) Myxobacter I and (iv) Myxobacter II. They are inhibited by chelating agents such as EDTA but not by sulfhydryl agents or DFP (Rao et al., 1998).

1.5.2. Applications of proteases

Proteases in addition to their highest commercial values play a crucial role in numerous pathological processes as well. Serine and metallo-proteases are the principal classes of proteases found in several species such as *Bacillus subtilis, B. amyloliquifaciens* and *Pseudomonas* sp. (Fugishige *et al*., 1992). Other potential industrial applications of alkaline protease include the utilization in peptide synthesis, in the resolution of the rasemic mixture of amino acids, in the hydrolysis of gelatine layers of X-ray films and the recovery of silver (Anwar and Saleemuddin, 1998; Kumar and Takagi, 1999).

Proteases in addition to their highest commercial values play a vital role in numerous pathological processes. Serine and metallo-proteases are the principal classes of proteases found in several species such as *Bacillus subtilis, B. amyloliquifaciens* and *Pseudomonas* sp. (Fugishige *et al*., 1992). Identification and characterization of microbial proteases are the prerequisite for understanding their role in pathogenesis of infectious diseases as well as to improve applications in biotechnology. For this, rapid and sensitive techniques for detection and characterization of microbial proteases are highly desirable (Lantz and Cibrowski, 1994).

A few proteases from extreme halophiles have been reported in some archaeal halophiles (Studdert *et al*., 2001), but few proteases from moderately halophilic and halo-tolerant bacteria have also been purified and studied (Sanchez-Porro *et al*., 2003). As moderate halophiles are capable to grow over a wide range of salt concentrations (optimal growth at 3–15% NaCl), they constitute a very interesting group of organisms with great potential for use in biotechnology because of their high activities at high salt concentrations (Ventosa *et al*., 1998).

1.5.3. Sources of proteases

Proteases are physiologically necessary for living organisms, they are ubiquitous in nature, found in a wide diversity of sources such as plants, animals and microorganisms (Rao et al., 1998). Enzymes can be separated from living cells and perform catalysis independent of their physiological environment. Commercial proteases are derived from animal tissues, plant cells and microbial cells via fermentation. Proteases of animal origin are secreted in the form of inactive precursor called pro enzyme or zymogen, in contrast nearly all of the microbial and plant proteases are produced in their active form.

1.5.4. Protease producing microbes

The inability of the plant and animal proteases to meet current world demands has led to an increased interest in microbial proteases (Rao et al., 1998). Proteases of bacteria, fungi and viruses are increasingly studied due to their importance and subsequent applications in industry and biotechnology. Commercial application of microbial

proteases is attractive due to the relative ease of large-scale production as compared to proteases from plant and animals. Also microbial proteases possess all characteristics desired for their biotechnological applications.

Microbial proteases account for approximately 40% of the total worldwide enzyme sales. (Rao et al., 1998). Microorganisms represent an attractive source of proteases as they can be cultured in large quantities in a relatively short time by established fermentation methods producing an abundant, regular supply of the desired product. Besides they can be genetically manipulated to generate new enzymes with altered properties that are desirable for their various applications (Rao et al., 1998, Gupta et al., 2002a). In general microbial proteases are extracellular in nature and are directly secreted into the fermentation broth by the producer, thus simplifying downstream processing of the enzyme in comparison to proteases obtained from plants and animals (Gupta et al., 2002a). Microbial proteases, especially from *Bacillus* sp. have traditionally held the predominant share of the industrial enzyme market of the worldwide enzyme sales with major application in detergent industry (Beg et al., 2003). Microbial enzymes are produced by growing the microbial cells under specialized conditions so that these cells produce their maximum amount of active enzymes. It is important to control environmental conditions during production so that a high percentage of these active catalysts are preserved intact (Shumi et al., 2004).

1.6. Extremophiles as a source for industrial enzymes

Life and growth of living organisms are monitored by numerous physical and chemical factors, each organism being defined and characterized according to specific parameters which are essential for its growth and development. Man thrives on the surface of the earth where temperatures are generally temperate, i.e., from 4 to 40°C, pH is around neutrality, or between 5 and 8.5, and where salinity, hydrostatic pressure and ionizing radiations are low. It is thus natural to consider that every ecological system that is not compatible with the growth and survival of humans as being “extreme”. Unlike many organisms that cannot survive outside of “normal” conditions, extremophilic microorganisms thrive optimally when one or several of these parameters are in the extreme range.

Extremophiles are bizarre organisms that can grow and thrive in extreme environments, which are formally considered too hostile to support life and the are rich source of varied critical materials (Table 1.2) which includes biocatalysts called extremozymes, produced by these microorganisms function under extreme conditions. Extremozymes offer new opertunities for biocatalysis and biotransformation due to their extreme stability. Currently, only 1 to 2 present microorganisms on the earth have been commercially exploited and amongst them only a few examples of extremophiles. However, the renewed interest currently emerging as a result of new developments in the cultivation and production of extremophiles will increase the biocatalytic applications of extremozymes (Joseph and Walter., 2004). The term extremophile was first used in 1974 (Mac Elroy, 1974). Extreme environment is a relative term, since environments that are

extreme for one organism may be essential for the survival of another organism. Extremophiles thrive under conditions that would kill most other creatures and many cannot survive in the normal anthropogenic global environments. Extreme environments include those with either high (50 to 121°C) or low (–2 to 20°C) temperatures, high salinity (2–5 M NaCl) and either high alkalinity (pH>8) or high acidity (pH<4) (Madigan and Marrs, 1997, Rothschild and Manicinelli, 2001).

Table 1.2. Biotechnological applications of Extremophiles.

Microorganisms	**Enzyme/ Endogene**	**Application/ Product/ Compound**
Thermophiles 50-121^{o}C	Amylases	Glucose, Fructose for sweeteners
	Xynalases	Paper Bleaching
	Proteases	Amino Acids, Food Processin
	DNA Polymerases	Baking, Brewing, Detergents Genetic Engineering
Psychrophiles 5-20^{o}C	Neutral Proteases	Cheese ,dairy protease production
	Amylases	Detergent
	Lipases	Pharmaceutical
	Poly saturated Fatty Acids	Artificial snow
	Ice Proteins	
Acidophiles pH<2	Sulphur oxidation	Mineral leaching ,coal desulphurization
Alkaliphiles pH>8	Protease, Amylase Lipase	Detergents
	Cyclo dextrins	Stabilization of Volatiles
	Antibiotics	Pharmaceuticals
Halophiles 3-9% salt	Carotene	Food cooking, glycerol,Pharmaceuticals compatible solvent, membrane surfactants for pharmaceuticals

The use of enzymes as biotransformation catalysts has been the subject of numerous texts and reviews. Mesophilic organisms are the major source of majority of enzymes used to date, despite their many advantages, the application of these enzymes is restricted because of their limited stability to extreme conditions of temperature, pH, ionic strength, etc. The majority of extremophiles are members of the Archaea, defined by comparison of 16S rRNA gene sequences, although extremophilic members of the bacterial domain are also known (Brown and Doolittle, 1997). In laboratory, microorganisms are generally cultivated as cell suspensions in liquid media. However, their natural environments are often more complex, and most microorganisms are members of mixed populations, growing frequently in a nearly static situation (e.g., embedded in soil or as biofilms) resulting from the passage of an aqueous stream of dissolved gases and nutrients over them. A wide variety of surfaces are available in natural environments for attachment and colonization by microorganisms, and these attached bacteria are often more active than free cells (Van de Loosdrecht et al., 1995).

1.7. Importance of extremophiles in protease production

Bacillus is a rod-shaped, gram positive, spore forming, aerobic, usually catalase positive, chemoorganotropic bacterium. Most bacillus species are harmless to humans and animals and have been used in several tradional food fermentations. Only a few pathogens are known, including *B. anthracis,* the causative agent of anthrax (Thorne, 1993), *B. cereus* which cause food poisioning (Aronson, 1993). The genus *Bacillus* is an important source of commercial enzymes, such as cellulases, lipases, starch degrading enzymes, and proteases (Ferrari et al., 1993). It may be concluded from the fact that this

genus has attractive properties for the production of industrially important enzymes (Park, 1996). Bacillus species have been major protease producers amongst industrial microorganisms having roles in applied microbiology (Schallmay et al., 2004).

They are active producers of extracellular proteases, and characteristics of enzyme production by bacillus species have been well studied. Bacillus species are grown on substrate media by including sporo genesis of bacteria which provide protease (Ayse et al., 2005).Alkaliphilic *Bacillus* can be found mostly in alkaline environments such as soda soils, soda lakes, neutral environments and deep-sea sediments. Animal manure, man-made alkaline environments such as effluents from food, textile, tannery, potato processing units, paper manufacturing units, calcium carbonate kilns and detergent industry are also good sources (Akbalık, 2004). A myriad of *Bacillus* species from many different exotic environments have been explored and exploited for alkaline protease production but most potential alkaline protease producing bacilli are strains of *B. licheniformis, B. subtilis, B. amyloliquefaciens,* and *B. majovensis* (Gupta et al.,2002a). Alkaline proteases produced by thermophilic and alkaliphilic bacilli can withstand high temperature, pH, chemical denaturing agents and in non-aqueous environments (Johnvesly and Naik, 2001).

Most of the studies have been performed to examine enzyme biotechnology, as alkaliphilic *Bacillus* strains produce enzymes, such as xylanases, cellulases, amylases, and proteases that are very useful in industry and domestic life (Kulkarni et al., 1999). Horikoshi reported the production of an extracellular alkaline serine protease from

alkaliphilic *Bacillus* strain 221 (Horikoshi, 1999). Moreover, various types of alkaline proteases have been characterized and their potential industrial applications have been explored. The spore-forming alkaliphilic *Bacillus* species constitute a large, diverse group of microbes (Takami and Horikoshi, 2000).

Metabolic activities of *Bacillus sp*. through the degradation of proteins, contribute directly or indirectly to the development of flavor and texture of fermented products. The proteolysis system of Bacillus may also contribute to the liberation of bio-active peptides which could play a significant role in enzymatic process involved during fermentation (Ouoba et al., 2005). Therefore, alkaliphilic Bacillus strains are quite important and interesting for both academic and industrial reasons (Denizci et al., 2004).

Studies have been conducted for the exploration of new type of alkaline proteases which are not only pH stable but also temperature stable in the presence of detergents. Proteases in addition to their highest commercial values play a crucial role in numerous pathological processes. Serine and metallo-proteases are the principal classes of proteases found in several species such as *Bacillus subtilis, B. amyloliquifaciens* and *Pseudomonas sp.* (Fugishige et al., 1992).

In the recent years, the use of alkaline proteases (that can retain stability at high alkaline pH values) in a variety of industrial processes like processing of food, leather, silk and detergents has increased remarkably (Gessesse, 1997). A large proportion of commercially available alkaline proteases are derived from *Bacillus* strains because of its ability to secrete large amounts of alkaline proteases having significant proteolytic

activities and stability at considerably high pH values and temperatures (Yang *et al*., 2000). Beside, the studies of Denizci *et al* (2004) and Horikoshi (1971) other studies have reported alkaline proteases from other strains of alkaliphilic *Bacillus* such as NKS-21 (Tsuchida *et al*., 1986), B-21-2 (Fujiwara and Yamamato, 1988) and *B. thermoruber* BT2 (Manachini *et al*., 1988). An alkaliphilic strain of *Streptomyces* species also produces alkaline protease (Nakanishi and Yamamoto, 1974).

Production of protease by microbial strains exhibits a characteristic relationship with regard to the growth phase of a specific organism. Protease synthesis in most of the *Bacillus* species is constitutive or partially inducible and is controlled by numerous complex mechanisms operative during the transition state between exponential growth and stationary phase. Hence, production levels are very much dependent upon the culture conditions (Strauh and Hach, 1993). In an alternative approach, alkaliphilic *Bacilli* that could grow at pH 9 and more were used as a source for oxidation-resistant alkaline proteases (Saeki *et al*., 2002). Within the past two decades on the classification of alkaliphilic *Bacillus* species according to phylogenetic and phenotyphic characterizations, described nine new species with the following names: *B. clausii*, *B. gibsonii*, *B. horikoshii*, *B. pseudoalkaliphilus*, *B. pseudofirmus*, *B. agaradhaerens*, *B. clarkii*, *B. halmapalus* and *B. halodurans* in addition to previously described species *B. alcalophilus* and *B. cohnii* (Neilsen *et al*., 1995).

Different assays have been developed for detection and presence of extra-cellular proteases from microorganisms on the agar plates in which different substrates like

gelatin, Bovine serum albumin (BSA), hemoglobin (Vermehlo et al. 1996). For all microorganisms under study, it was found that in agar-BHI (Brain Heart Infusion) or yeast extract medium containing gelatin, the sensitivity of protease detection was considerably greater than in BSA-agar or hemoglobin agar. However, when BSA or hemoglobin was added to the culture medium, there was an increase in growth along with a marked reduction in amount of protease production. In these studies proteolytic activity was attributed by the size of the clearing zone on the plates. Activity of protease enzyme in the fermented medium is assessed by measuring the release of trichloroacetic acid soluble peptides from different substrates(Dapeau, 1976). Most widely used substrates for the purpose are the casein and azo-casein for their easy availability and being cheap. One of the amino acids found upon hydrolysis is L-tyrosine. This product strongly absorbs light at 280 nm and concentration of tyrosine is determined as a function of optical density in accordance with Beer's law (Dapeau, 1976).

Same procedure and substrate (0.6% Hammerstein casein) were used for the determination of alkaline protease activity in strain of *Bacillus clausii* (GMBAE M2) that was found to be capable of growing under highly alkaline conditions (Denizci *et al*., 2004). One unit of alkaline protease activity was defined as the amount of enzyme able to produce 1µg tyrosine in 1 minute under assay conditions. *Bacillus sp*. grows very fast and formation of protease starts after 5 h of growth and reaches its maximum in 9 hours (1.93 U / mg proteins) and then begins to fall (Ward, 1985). This suggests that protease production is directly linked to culture being metabolically active. *Bacillus sp*. usually secretes more protease in late exponential phase. The function of this enzyme is obscure

but its production is correlated with the onset of high rate of protein turnover during sporulation (Ward, 1985). Several reports have indicated that protease production is dependent on environmental conditions. Extracellular proteases have been shown to be sensitive to repression by different carbohydrate and nitrogen sources (Haulon *et al*., 1982). Other factors may have effect on the growth phase and pH (McIver and Scott, 1997). The extracellular proteases of the genus *Bacillus* are mainly either serine (alkaline) or metal (neutral) enzymes (Priest, 1977).

In the recent past, a renewed interest has been generated on the utilization of alkaline proteases produced by alkalophilic microorganisms (Horikoshi, 1999). *Bacillus* derived Moreover, alkaline proteases specifically account for nearly 25% of the world enzyme market with a predominant share (35%) taken by detergents (Layman, 1986). In general, the majority of the available alkaline proteases for industrial use face some limitations such as low activity and stability towards anionic surfactants like sodium dodecyl sulphate (SDS) and oxidants like peroxides and perborates, which have been the common ingredients in modern bleach-based detergent formulations. Further, about 30–40% of the production cost of industrial enzymes is accounted for by the cost of the growth medium. Considering these critical factors, Joo et al (2003) attempted the isolation and screening for novel marine extremophiles and employed some low cost and easily available medium ingredients for cost effective alkaline protease production. In a later study, *Bacillus clausii* I-52 isolated from the heavily polluted tidal mud flat was reported to secrete a strong alkaline proteolytic enzyme having industrial potential (Joo and Chang, 2005).

However, other application potentials of these enzymes depend on the nature of catalytic activity with respect to reactant medium, which led to the classification of proteases as acidic, neutral, and alkaline. Among these different biocatalysts, alkaline proteases have wide application spectra and novel properties due to their exotic catalytic nature. Hence, these proteases and their producing organisms attracted attention of scientific community to understand the protein chemistry and protein engineering to enhance their utilization niche (Germano *et al*., 2003). Though several life forms such as bacteria, fungi, yeast, plant, and mammalian tissues are known to produce alkaline proteases (Prakasham *et al*., 2005), with increasing industrial demand for proteases it is expected that hyperactive strains will emerge and that the enzymes produced by new exotic microbial strains could be used as biocatalysts in the presently growing biotechnological era. As a conclusion and available literature information indicate that, among all protease-producing microbial organisms, the *Bacillus* genus has assumed importance because of its extended potential for production in large amounts (Montville, 1983).

1.8. Metagenomic approaches to obtain proteases

The rapidly emerging field of metagenomics seeks to examine the genomic content of communities of organisms to understand their roles and interactions in an ecosystem. (Yooseph et al., 2007) The inability to culture most microorganisms from environmental samples is a fundamental obstacle to understanding microbial ecology and diversity (Yeates et al.1998). The use of DNA-based techniques can overcome this

limitation by allowing the fate of particular genes or organisms to be monitored directly in environmental samples. Isolation of bacterial nucleic acids from natural environments has become a useful tool to detect bacteria that cannot be cultured to determine the fates of selected bacteria or recombinant genes under natural conditions and to reveal genotypic diversity and its change in microbial ecosystems (Zhou et al ., 1996).

Culture dependent methods rely on the chemical composition of microorganisms in the environment. Bacteria change their chemical composition substantially to accommodate environmental fluctuations. Thus, when comparing bacteria on the basis of some chemical components, it is important that variation observed is a result of genetic differences and not due to an environmental effect. Cultures must, therefore, be grown under identical conditions and to the some stage of the batch culture growth cycle to ensure uniformity of the environmental influence. This may be particularly difficult, some time impossible, if physiologically diverse organisms such as thermophiles and psychrophiles or aerobes and microaerophilic taxa are being compared.

Of the various components used for taxonomy only chromosomal DNA and RNA are unaffected by growth conditions. The amount of these molecules will fluctuate with growth rate, but the composition is invariant. Thus the nucleic acids offer the only standard molecules by which the widest range of microorganisms (and higher eukaryotes) can be compared and classified. Constitutively synthesized proteins are also very useful in this respect but individual proteins may not be distributed universally (Brian and Fergus 1986).

Developing new strategies to recover the microbial population that is not accessible by using classical enrichment techniques has become a real issue for microbiologists. Molecular studies of microbial diversity have shown that these “not yet cultivated” microorganisms can correspond to 99% of the total microbial population (Amann *et al*., 1995). Since all drugs isolated from microbes have only been made from cultivatable microbes, it is clear that the rest, i.e., 99% of bacteria and 95% of the fungi could be a major source of future pharmaceuticals. This requires the study of ‘‘environmental DNA’’. Metagenomic examination of the ‘‘uncultivated majority’’ is being carried out by development of environmental gene tags, serial analysis of 16S rRNA gene sequence tags, new screening techniques for metagenomic libraries, and new methods of cultivation(Greden, B. D., & Keller, M. (2006)

Genetic fingerprinting techniques provide a pattern or profile of the genetic diversity in a microbial community. Specific DNA sequences can give important information about the community structure and diversity of microbes containing a critical gene. These techniques are important in separating and identifying PCR-amplified products that might have the same size but slightly different nucleotide sequences. In the past two decades, these molecular tools exemplified by 16S rRNA analysis have facilitated the study of natural microbial populations without cultivation which has made quantitative assessment of microbial diversity now conceivable. Many variations of the 16S rRNA approach are currently used for defining microbial diversity. These include analysis of PCR amplified 16S rDNA sequences and their digestion with restriction

endonucleases to obtain RFLP of whole 16S rDNA . However, no single tool allows definite assessment of the bacterial diversity. Therefore, the use of a combination of molecular biology techniques, microbiological methods and geochemical techniques or microsensors (Ramsing *et al*., 1996; Teske *et al*., 1996) is necessary to obtain a better understanding of the interaction between the microorganisms and their natural environment.

The primary structure of 16S rRNA contain stretches of sequence conserved to varying degree and their positions are mostly known. Sequence information from the conserved regions is useful for studying phylogentic relationships (Woese *et al*., 1985) as well as for the design of universal oligonucleotide probes and primers used for identification and amplification, respectively (Giovannoni, 1991). Variable regions provide sequence data to develop specific probes and primers for detection of microorganisms by hybridization or with polymerase chain reaction (Ward *et al*., 1992). Using conserved primers, the16S rRNA gene can be easily amplified by PCR not only from pure cultures but also directly from the environmental samples (Giovannoni *et al*., 1990 and Olsen *et al*., 1986).

1.9. Soil metagenomics

Soil microorganisms represent a significant portion of living matter on Earth and play a key role in ecosystem functions. Bacteria, fungi, archaea and viruses are the four microbial groups currently known to man. Bacterial presence in soil has been the most extensively studied, however with the environment at the forefront of worldwide focus,

expanded research on fungal, archaeal, and viral communities is much needed (Fierer et al., 2007).

In the laboratory, microorganisms are generally cultivated as cell suspensions in liquid media. However, their natural environments are often more complex, and most microorganisms are members of mixed populations, growing frequently in a nearly static situation (e.g., embedded in soil or as biofilms) resulting from the passage of an aqueous stream of dissolved gases and nutrients over them (Nathalie et al., 2003). A wide variety of surfaces are available in natural environments for attachment and colonization by microorganisms, and these attached bacteria are often more active than free cells (Van et al., 1990). Furthermore, access to this reservoir of genetic and metabolic diversity is of interest in areas such as food production, medicine, bioremediation of waste materials, and agriculture. In particular, there is an important biotechnological interest in the isolation of extremophilic microorganisms that are adapted to growth under extreme conditions of temperature, pH, salinity, pressure, and/or in the presence of radioactivity or high concentrations of metal. These extremophiles present a commercial potential for various industries and products, including agricultural applications, chemical synthesis, laundry detergents, and pharmaceuticals (Rothschild and Mancinelli., 2001).

1.10. Proteases in soil

Organic nitrogen applied to soil undergoes a stepwise mineralization process including proteolysis, ammonification, and nitrification. Proteases mediate the first steps of mineralization, which often are rate-limiting for the nitrogen cycle (Ladd and Butler, 1972). Soil proteases are derived mainly from heterotrophic soil bacteria (Asmar et al.,

1992; Nannipieri et al., 1983). These proteolytic bacteria include *Pseudomonas fluorescens, Bacillus cereus, B. mycoides,* and *Flavobacterium-Cytophaga spp.* in a broad range of soil types, and appear to secrete mostly metalloproteases (Bach and Munch, 2000,2001). For example, Hayano et al. (1987) reported that metalloproteinases were the dominating proteases in extracts from an andosol collected from a tomato field.

Different proteolytic bacteria release different amounts or activities of protease, and that the composition of proteolytic bacterial communities may play a major role in determining overall soil protease activity. (Sakurai et al., 2007).There is evidence that the level of soil proteolytic activity could be rate-limiting in the mineralization of soil organic nitrogen .When growth of microorganisms is retarded (Asmar et al. 1992a). Soil proteolytic activity can be influenced by various factors such as organic carbon concentration, total nitrogen concentration, clay content or total exchangeable cations (Ladd and Butler 1972) as well as the amendment of energy-rich substrates (Asmar et al. 1992a). Only limited research has been conducted to elucidate the microbial sources of soil proteolytic enzymes which would enable us to investigate the influences to the formation of peptidases more precisely. By selective inhibition of specific microorganisms it has been shown that soil peptidases are mainly of bacterial origin in a paddy soil (Hayano and Watanabe 1998). Asmar (1992b) showed a high correlation between proteolytic activities and total counts of bacteria in glucose-amended soils. DNA can be isolated from bacterial fractions containing 50 to 80% of the soil bacteria and may provide genetic information about the nonculturable bacteria in soil. The heterogeneity of

this DNA is a measure of the total number of genetically different bacteria in soil (Torsvik et. al., 1990)

Soils proteases are liberated by the microbial community and hydrolyze proteins to amino acids. Hydrolysis, like ammonification and nitrification, is an important process in the nitrogen cycle which plays an essential role in soil fertility. Information on the properties of the proteases is required before the N transformations in soil can be understood. However, there is little information about the properties of various soil proteases (hayyano 1987).Soil protease activity has been correlated with the number of soil bacteria. The sources of soil enzymes have been determined by monitoring the decrease in enzyme activity after selective inhibition of specific microorganisms (Watanabe 1994).

Microbial activity has been positively correlated with the activity of enzymes such as dehydrogenase, protease, cellulase, phosphatase, and urease in many soils, both in the field and in the laboratory. However, the distinction between intracellular and extracellular enzymes in soil assays has certain shortcomings, because there is no ideal method that can discriminate between extra- and intracellular enzymes.(Asmar et al ., 1992) Little information is available about how bacterial growth induced by root released organic C and N in the rhizosphere and successive protozoan grazing affects the activity of soil enzymes involved in the ammonification process, such as proteases and deaminases(Badalucco et al ., 1996). Here it can be concluded that soil microbes are a

rich source of protease production and metagenomic techniques can help to explore noval protease genes present in these microbes.

1.11. Genomic studies on proteases

Genetic studies on extremozymes have received considerable interest lately, propelled by developments in microbial isolation and the application of genomic approaches for studying the individual organisms and their community (Valenzuela et al., 2006). Among the alkaliphilic microorganisms exhibiting protease activity the members of the genus *Bacillus* were found to be abundantly present (Kumar and Takagi, 1999). Therefore, alkaliphilic *Bacillus* strains are quite important and interesting both for academic and industrial reasons. A significant number of genomes from extremophilic bacteria have and are currently being elucidated. Large potential of extremophiles as a source for novel enzymes is the observation that was made after determining the 1.56 Mb genome of *Thermoplasma acidophilum* (Ruepp *et al*., 2000). Genetic tools for *E. coli* expression system are very well developed and many plasmid vectors and inducible gene expression systems are available. In particular, the production of thermophilic enzymes is facilitated by the efficient denaturation of most *E. coli* proteins following a short heat treatment (Imanaka *et al*., 2002).

However, some heterogeneity in both phenotypic and genetic characteristics has been found in strains putatively aligned with given phena, particularly strains assigned to the phenon encompassing the species *Bacillus clausii* (Nielsen *et al*., 1995), which indicates that there are intrinsic difficulties in identifying the alkaliphilic strains at the

species level. Neutral and alkaline proteases from *Bacillus subtilis* and *Bacillus amyloliquefaciens* have been cloned and sequenced (Wong *et al*., 1984). The genome of a haloalkaliphilic archaeon from highly alkaline soda lakes, *Natronomonas pharaonis*, has revealed several adaptations to this environment (Falb *et al*., 2005). These include an overall modification of the proteome to increase the fraction of acidic amino acids and reduce protein hydrophobicity, a coating of the cell membrane with glycoproteins and secreted enzymes attached by lipid anchors, and an efficient transport system for heavy metals and nitrogen compounds which are scarce in hypersaline environments. The halophilic bacterium, *Salinibacter ruber* (Mongodin *et al*., 2005) displays similar adaptation mechanisms to hypersaline environments and some of these could have been acquired via lateral gene transfer. No doubt many more surprises are hidden in these genomes that can be exploited for biotechnological purposes.

A wide variety of *Bacillus* species (Markland and Smith, 1971) secrete serine endoproteases into the external medium. *Bacillus* serine proteases have their best-known application in detergent powders. To best meet the alkaline conditions in detergents, serine proteases with a highly alkaline pH optimum (referred to herein as high-alkaline proteases) are preferred above the subtilisins that have an optimal pH of 8.5 to 10. After a screening programe, Zuidweg *et al*. (1972) isolated an alkalophilic *Bacillus* strain (PB92) that produced a high-alkaline serine protease (PB92 protease) with unique pH optimum of 10.5 to 12. The PB92 protease performed extremely well in detergents. Amino acid sequences for the serine proteases from *Bacillus amyloliquefaciens*, *Bacillus licheniformis*, and *Bacillus subtilis* have been determined (Nedkov *et al*., 1985), and the

genes from *B. amyloliquefaciens*, *B. licheniformis*, *B. subtilis*, *B. subtilis subsp. amylosacchariticus*, and the alkalophilic *Bacillus sp*. strain YaB have been cloned and sequenced (Kaneko *et al*., 1989).

Many species of the genus *Bacillus* produce a variety of extracellular enzymes (Priest, 1977), some of which are of industrial importance. Studies on these extracellular enzymes are significant not only from a practical standpoint, but also from a theoretical point of view for better understanding the secretion mechanism. The molecular cloning of the structural genes for such enzymes, e.g., penicillinase of *B. licheniformis* (Imanaka *et al*., 1981) and amylase of *B. amyloliquefaciens* (Palva, 1982) in *Bacillus* species, has conspicuously enhanced the productivity of these enzymes and provided a closer look at the secretion of each enzyme across the cell membrane. When the signal peptides of these extracellular enzymes are identified, the information will be most useful for the efficient secretion of a foreign gene product as a fused protein (Chang *et al*., 1982).

The DNA sequences of the cloned subtilisins reveal an intervening propeptide sequence between a signal sequence and the mature enzyme (Jacobs *et al*., 1985). It was shown by analysis of sub cellular fractions of *Bacillus* strains expressing various mutations in the subtilisin gene that the proprotein of subtilisin exists in association with the cell membrane. Furthermore, the conversion of the primary gene product into the mature enzyme is mediated by active subtilisin, and therefore this processing is most likely autocatalytic (Powers *et al*., 1986).

It is generally true that the enzymes of an organism are adapted to function optimally at or near its growth conditions, and so the range of extremes at which life is found defines the range of conditions at which enzyme activity might be detected. The greater the diversity of organisms detected in the environment, the greater the potential range of biocatalysts available for applications such as biotransformations. The classical route to microbial enzyme production, involving isolation of the organism as a pure culture followed by purification of the enzyme either from its natural host or by cloning and expression of the corresponding gene, has been used to obtain extremozymes (Connaris *et al*., 1998). Alternative approaches have been applied in situations where pure cultures are not available, however. These include screening of expression libraries prepared from environmental DNA (Winson and Kell, 1997) or cloning of PCR-amplified gene fragments to produce hybrid genes for expression in a mesophilic host (Okuta *et al*., 1998).

To overproduce a protease, two strategies for increasing the gene copy number in *Bacillus* can be considered: (i) using plasmid-containing strains and (ii) using strains containing additional genes integrated in the chromosome. In *B. subtilis*, replicative plasmids harboring cloned sequences can be prone to either segregational or structural instability (Harington *et al*., 1988). Chromosomal integration has been applied to solve these problems. Although these strains are more stable than plasmid-containing strains, instability of tandemly amplified chromosomal sequences has also been reported (Albertini and Galizzi, 1985).

To maximize industrial enzyme production, the host organism is generally manipulated to carry multiple copies of the gene. This can be accomplished by cloning the gene on a replicating plasmid, but preferentially this is achieved by amplification of the gene on the chromosome, as such amplification offers the more stable alternative. Microorganisms such as *Bacillus subtilis*, which can easily be transformed and has well-developed genetics, it is relatively straightforward to amplify genes on a chromosome. However, other *Bacillus* species frequently used in industry are more difficult to transform and less amenable to genetic manipulation.

It is possible to use temperature-sensitive plasmids, such as pE194, as a means of achieving gene amplification on a chromosome; however, the incorporation of the plasmid origin of replication in the host chromosome is known to destabilize amplified structures (Young and Ehrlich, 1989). There is therefore an interest in the development of improved amplification methods for industrial *Bacilli*. Possibly the most significant recent advances in molecular microbiology have resulted from rapid progress in the field of genomics.

The NCBI (http://ww.ncbi.nlm.nih.gov) Genome database provides views for a variety of genomes, complete chromosomes, sequence maps with contigs, and integrated genetic and physical maps. The database is organized in six major organism groups: Archaea, Bacteria, Eukaryotae, Viruses, Viroids, and Plasmids and includes complete chromosomes, organelles and plasmids as well as draft genome assemblies (Fig.3)

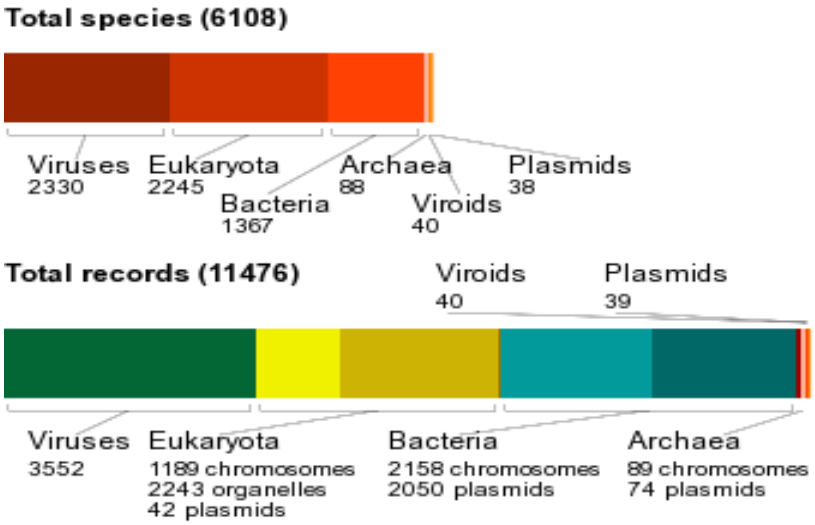


Fig.3: Overview of genome records for different microorganism at NCBI.

Moreover, the dramatic advances in molecular and computational biology, new sources of enzymes, combinatorial methodologies and biochemical engineering of single and multicomponent enzyme systems have stimulated a renaissance in enzyme technology. Revolutionary changes in molecular biology are leading to the development of thousands of new uses of enzyme technology, fueling major growth in this multi-billion-dollar field. A huge amount of money is being invested worldwide in the development of industrial as well as biomedical applications of extremozymes and extremophiles (Oren, 2002). As we learn more about the extremes at which life can survive and thrive, more of these extremophiles are brought into culture and their genomes sequenced. In many cases, biotechnological applications of extremophiles and their biomolecules (e.g. enzymes) have been the driving force in both academic and industrial research of these organisms. Extremophiles and extremozymes occupy an important place in the multibillion dollar environmental biotechnology industry, with

applications spanning agricultural, biomedical and industrial sectors (Egorova K and Antranikian, 2005).

Novel developments in the cultivation and production of extremophiles, but also developments related to the cloning and expression of their genes in heterologous hosts, will increase the number of enzyme-driven transformations in chemical, food, pharmaceutical and other industrial applications (Van den Burg, 2003).

1.12. Use of bioinformatics in gene and protein analysis

The acquisition and management of large scale data sets are driving innovations in biotechnology and bioinformatics. Bioinformatics deals with the research and management of genomic and proteomic datasets. It gives the computational prowess as well as mathematical insight to genetic information and captures their biological significance through insilico means. It also provides the means to interpret the data as well as to predict noval results based on existing genetic and genomic information. There are number of software available, also a aplenty of web based analysis tools, for sequence analysis , structure analysis and high through put array research.

1.12.1. Applications of bioinformatics

Bioinformatics is the use of IT in biotechnology for the data storage, data ware housing and analyzing the DNA sequences. In Bioinfomatics knowledge of many branches are required like biology, mathematics, computer science, laws of physics & chemistry, and of course sound knowledge of IT to analyze biotech data. Bioinformatics is not limited to the computing data, but in reality it can be used to solve many biological

problems and find out how living things works.It is the comprehensive application of mathematics (e.g., probability and statistics), science (e.g., biochemistry), and a core set of problem-solving methods (e.g., computer algorithms) to the understanding of living systems.

Bioinformatics is being used in following fields: Molecular medicine, Personalised medicine, Preventative medicine, Gene therapy, Drug development, Microbial genome applications , Waste cleanup, Climate change Studies, Alternative energy source, Biotechnology, Antibiotic resistance, Forensic analysis of microbes, Bio-weapon creation, Evolutionary studies, Crop improvement, Insect resistance, Improve nutritional quality , Development of Drought resistance varieties, Vetinary Science and many more.

It was through the combined efforts of using information technology and computer science that allowed them to create a large database capable of housing and securely storing all of the valuable work that was being done with studies on DNA. The database that was created allowed scientists to be able to access millions of records of parts and molecules of DNA sequences from different species to compare to the work that was currently being done.

When it comes to the different applications of bioinformatics, the options are limitless. There are constantly being new research projects and studies being done on this amazing new line of DNA analysis. Scientists are now using bioinformatics to detect genetic abnormalities in different species. This is also creating breakthroughs in the medical community. Applications of bioinformatics have allowed doctors to conduct

genetic testing in unborn babies to predict and find any signs of certain genetic disorders and conditions. While some of the work that bioinformatics has created is under scrutiny, other areas are taking off in the world of science and discovery. The doors are opening for scientists to be able to find and predict genetic problems before they exist. The world of bioinformatics is finding new applications at all times.

1.12.2. Various tools used in bioinformatics studies

There are various tools available both in the form of softwares as well as web based analysis programs for the analysis of biological data using computational knowledge.

1.12.2.1. DNA and protein sequence analysis

For the determination of sequence similarities, pairwise sequence alignment of DNA and proteins is carried out as a preliminary step. For this purpose BLAST (McGinnis et al., 2004) and FASTA (Silvanovich et al., 2009) are well known programs. BLAST (Basic Local Alignment Search Tool) is widely used program for local alignment available at NCBI (http://www.ncbi.nlm.nih.gov) FASTA (Silvanovich et al., 2009) is global alignment tool.

1.12.2.2. Multiple sequence analysis

Multiple sequence alignment is useful for the analysis of DNA level information to seek clues to polymorphism affecting populations as well as halotypes, that have a protective effect on a population. On the other hand it is used to understand proteins, protein families, and superfamilies. It is also used to find conserved motifs in proteins that are critical in physiological function and finally finding the ancestral regions of

proteins, or evaluation of the evolutionary characteristics of a set of protein. A most frequently used program for this purpose is Clustal X (Thompson et al., 1997)

1.12.2.3. Phylogenetic analysis

Phylogenetic methods can be used for many purposes, including analysis of morphological and several kinds of molecular data with particular emphasis on DNA and protein sequences. phylogenetic analysis helps in comparisons of more than two sequences, analysis of gene families, including functional predictions, estimation of evolutionary relationships among organisms. There are many softwares available for this purpose among them PHYLIP (Felsenstein ., 1992)program is much used.

1.12.2.4. Sequence interpretation

For the interpretation of DNA and protein sequences various parameters can be applied like to find out open reading frame in a DNA sequence ORF finder is commonly used program. Restriction site analysis can also be done using various web based also through different software. Likewise for protein sequence analysis many programs are designed to plot hydrophobicity profile, predict proteins secondary stucuture, motif and active site searching etc.There are many softwares available which deals with the interpretation of both DNA and protein sequences at a time.

1.12.2.5. Homology modeling

Three dimensional comparative models of different target proteins based on structures of template proteins can be predicted using different methods for this purpose various commonly known tools like SWISS MODEL (Guex et al.,1997), WHATIF(Gret

Vriend ., 1990), and MODWEB (Eswar et al., 2003) ,MODELLER (Eswar et al., 2003) etc are available.

1.12.2.6. Structure analysis

The quality and precision of the predicted 3D protein model can be evaluating using various visualization tools like VMD (Humphrey et al.,1996),RasMol(Bourne and Weissig, 2003), DS Viewer (Pazel., 1989) etc. for geometry analysis Procheck (Laskowski et al., 1996) and What check (Gret Vriend, 1990) are extensively used programs.

1.13. Aims of the present study

Our studies deal with the study of biodiversity contained in indigenous soil of varied nature for the exploration of industrially important protease gene. Moreover, protease genes have been PCR amplified, cloned and sequenced from various cultured microorganisms and metagenomic samples, to study molecular aspects of these genes using different molecular biology techniqures. Various bioinformatics tools have been employed for the characterization of these genes. Also protein structures prediction was performed to know the structural and functional features of protease gene.

CHAPTER

2.0

MATERIAL & METHODS

2.1. MICROBIOLOGICAL STUDIES FOR PROTEASE GENE

In order to explore protease genes in bacterial communities of diverse origin , bacterial samples were obtained from three different sources:

1. Environmental samples from pesticide factory
2. Bacterial isolates from NIBGE culture collection
3. Soil samples from NIBGE for metagenomic studies

2.1.1. Environmental samples from pesticide factory

Soil samples were collected in sterilized glass bottles from the sale point for insecticide as well as from different locations of canal, receiving drainage water of a factory involved in the formulation of pesticides.

2.1.1.1. Isolation of environmental samples

Soil samples were processed on the day of collection by mixing 10 grams of each soil sample in 100 ml autoclaved water and keeping them in orbital shaker (Kuhner Shaker) at 100 rpm for 24 hours at room temperature (30°C). From each processed sample, 0.5 ml was spread over the surface of nutrient agar media plates.

2.1.1.2. Preparation of media to obtain pure cultures

Nutrient agar being a rich medium, was used as general purpose medium. Its composition included 1% tryptone, 0.5% yeast extract, 1% NaCl and 1.5% agar dissolved in distilled water. Media were autoclaved at 121°C for 20 mins at 15 psi. The pH of media was adjusted to 7.2 before autoclaving. Molten media were poured into glass Petri plates and allowed to solidify. After solidification, plates were kept in oven for 24 hours to check any contamination in media. Cultures were streaked aseptically on solid media plates and incubated at 37°C till growth was observed. Separate colonies of each strain

were inoculated into the respective liquid media and growth was confirmed by the appearance of turbidity. Six pure cultures so obtained were preserved as nutrient agar slants in glycerol.

2.1.1.3. Colony and cell morphology studies

Apparently different colonies were grown on different overlay plates and were examined by stereomicroscope to study morphological features like colony color, colony form, colony elevation, and colony margins, etc.

2.1.1.3.1. Phase contrast microscopy

Phase contrast microscope Zeiss Axiovert model MC80 was used to record cell size, shape, morphology, motility and presence or absence of endospores etc. Microscope was fitted with a camera to take photographs of peculiar structures.

2.1.1.3.2. Gram's staining

Following solutions were prepared for Gram's staining.

A. Hucker's crystal violet reagent

B. Stabilised Lugol-PVP complex

C. Counterstain

A. Hucker's crystal violet reagent

Hucker's crystal violet reagent was made by mixing two solutions, solution A and solution B. Solution A was prepared by mixing 2.0g of crystal violet in 20.0ml of ethanol (95%). Solution B was prepared by mixing 0.8g of Ammoniumoxalate in 80.0ml of distilled water. Then these two solutions were mixed together.

B. Stabilized Lugol-PVP complex

This solution was made by mixing 1.3g of iodine, 1.0g of polyvinypyrroliodine and 2.0g of KI in 100.0ml of distilled water.

C. Counterstain

This solution was prepared by mixing 0.25g safranin and 10.0ml of ethanol (95%) in 100.0ml of distilled water.

Procedure

1. A thin film from fresh bacterial culture was spread on a sterile glass slide and was fixed by gentle heating on a Bunsen flame.
2. Then slide was flooded with Hucker's reagent for 1 minute.
3. After that slide was washed by dipping the slide into slow running tap water.
4. Then the slide was flooded with iodine solution for 1 minute.
5. The slide was placed diagonally in a glass box and iodine solution was rinsed off with safranin.
6. Then excess amount of fresh safranin was added for 35 seconds.
7. The slide was washed with water as described under (3).
8. Allowed the slide to air-dry.
9. Gram-positive cells appeared purple and Gram-negative cells pink.

2.1.1.4. Molecular characterization of isolates from environmental samples

2.1.1.4.1. DNA isolation

1. Bacterial cultures were centrifuged for 10 minutes at 10,000rpm to get cell pellet. 10-20mg cell pellet/tube was taken in several 1.5ml eppendorf tubes. 500µl TE buffer and 200µl of lysozyme (15mg/ml of TE buffer) was added in each tube. Then these were incubated at 37˚C for 24 hours.
2. 50µl RNase (10mg/ml) was added in each eppendorf tube and mixed gently by inverting the tube for several times. After mixing, tubes were incubated again at 37˚C for 24 hours.
3. 30µl SDS (10%) was added in each tube and incubated them at 70˚C for 30 minutes. Then 10µl Proteinase K (20mg/ml) was added in each tube and incubated again at 37˚C for 24 hours.
4. Tubes were centrifuged at 10,000 rpm for 10 minutes to remove cell debris and the supernatant was collected in fresh tubes.
5. 690µl chloroform: isoamyl alcohol (24:1) was added in the supernatant and mixed gently by inverting the tubes to form an emulsion.
6. The tubes were centrifuged at 13,000 rpm for 15 minutes.
7. The aqueous layer was taken in a new tube and discarded the remaining chloroform phase.
8. Equal volumes of phenol: chloroform: isoamyl alcohol (25:24:1) was added in each tube and centrifuged for 15 minutes. The aqueous layer was taken in fresh tube and 70µl 3M Na-acetate and 400µl of isopropanol was added in it.
9. Each tube allowed to stay at –20˚C for 1 hour.

10. Centrifuged at 14,000 rpm for 10 minutes and discarded the supernatant. Washed the DNA pellet with 70% ethanol to remove the bases.
11. Suspended the DNA pellet in appropriate volume of deionized H_2O.
12. The quality of DNA was checked by running on 1% agarose gel.

2.1.1.4.2. Agarose gel (1%) for DNA electrophoresis

1. Adequate volume of TAE buffer (1.0X) was prepared to fill the electrophoresis tank and to prepare the agarose gel.
2. For 1% agarose gel, desired amount of agarose was taken in a flask containing TAE buffer (1.0X) and was melted in microwave oven.
3. Melted agarose was cooled to 45-55°C and poured into the gel-casting tray. A comb was inserted into it.
4. Air bubbles were removed, if any, underneath the comb or on the surface of gel and the gel was allowed to solidify at room temperature.
5. After solidification of gel, the comb was removed carefully to avoid tearing of wells.
6. The gel casting tray, containing gel, was placed in electrophoretic tank, having 1.0X TAE buffer in it. Ethidium bromide (2 µl) was added to the tank.
7. DNA samples were mixed with appropriate volume of loading dye and loaded into the wells with a micropipette.
8. Voltage was set at 70 V. Movement of dye indicated the migration of DNA from anode to cathode through gel.
9. When dye covered the distance sufficient for separation of DNA fragments, the power supply was turned off.
10. DNA fragments were visualized under UV light and photographed.

2.1.1.4.3. PCR amplification of 16S rDNA

Polymerase chain reaction (PCR) amplification of 16S ribosomal DNA of isolates was carried out using the following primers:

Forward primer (FD1) AGAGTTTGATCCTGGCTCAG

Reverse primer (rP1) ACGG (ACT) TACCTTGTTACGACTT

50µl of reaction mixture was prepared for each amplification. 1X PCR buffer, $MgCl_2$, dNTPs and taq DNA polymerase were of Fermentas Company, while FD1 and rP1 primers were reported by Ghauri et al., 2003.

For 50µl PCR reaction mixture:

Nanopure water	34.75 µl
1X PCR buffer	5 µl
$MgCl_2$ (25mM)	6 µl
dNTPs (20mM)	1 µl
FD1	1 µl
rP1	1 µl
Taq DNA polymerase (1u/ µl)	0.25-1 µl
Template	1 µl

2.1.1.4.4. PCR profile for 16S rDNA amplification

Initial denaturation was carried out at 95 ^{o}C for 5 minutes. Amplification was carried out in thirty cycles. Each cycle was comprised of 30 seconds at 95^{o}C, 40 seconds at 55 ^{o}C and 2 minutes at 72 ^{o}C. The final extension was for 10 minutes at 72 ^{o}C.

Amplified PCR product was run on 1.5% agarose gel inadequate volume of 1.0X TAE buffer.

2.1.1.4.5. Phylogenetic analysis of isolates

Amplified 16S rDNA of about 1,500 bp from respective isolates was sequenced commercially (Macrogen Inc., Seoul, Korea). The gene sequences were compared with others in the GenBank databases using the NCBI BLAST (http://www.ncbi.mlm.nih.gov). Gene sequences of 16S rDNA of selected organisms were obtained from GenBank and aligned with gene sequence of our isolates using CLUSTALX. The aligned sequences were used to construct a distance matrix (Jukes and Cantor 1969), after the generation of 100 bootstrap sets, that was subsequently used to construct a phylogenetic tree by neighbor-joining method (Saitou and Nei 1987), using Treecon software (Van de Peer and De Wachter 1993).

2.1.1.5. Dual substrate plate diffusion assay for protease activity

It was done by following method presented by Montville (1983). Protease assay plates containing 1% casein, 1% gelatin in a solution of 1.5% agar were prepared. Because of direct addition of casein to the medium resulted in macroscopic aggregates and plates with poor readability, the casein was dissolved in 0.02N NaOH and stirred until it formed a translucent solution. The other compounds were added and the pH of the medium was adjusted to 9.0 with 1N HCl. The medium was sterilized at 121°C and tempered at 60°C and dispensed in 20 ml portions into sterile petri dishes. After the agar solidified, two wells were cut into the plates with a sterile cork borer for each sample, in one well 24 hrs old bacterial cultures supernatant from fermentation medium was poured

while in the second supernatant from control medium was poured. Petri dishes were incubated at 37°C. After 48 hrs, plates were stained for 15 minutes with commassie blue R-250 staining solution and destained with destaining solution containing methanol, acetic acid and water (in ratio 9:2:9 V/V/V). Protease activity was observed by the formation of a hollow zone around the well.

2.1.1.5.1 Protease activity of isolates obtained from environmental samples

The organism identified through 16S rDNA analysis were tested for protease activity but no significant results were found. That is why they were not considered for protease gene hunting.

2.1.2. Bacterial isolates from NIBGE culture collection

Six pure extremophilic bacterial isolates i.,e *Bacillus amyloliquefaciens* (BPT-05), *Bacillus cereus* (BPT-06), *Staphylococcus arlettae* (BPT-08), *Bacillus licheniformis* (BPT-20), *Bacillus licheniformis* (BPT-23) and *Bacillus pumilus* (BPT-25) were obtained from NIBGE culture collection and were originally isolated from environmental samples collected from Khewra salt mines (Nasrin et al., 2008).

2.1.2.1. Protease activity of isolates from NIBGE culture collection

All the six pure extremophilic Bacterial isolates i.,e *Bacillus amyloliquefaciens* (BPT- 05), *Bacillus cereus* (BPT-06), *Staphylococcus arlettae* (BPT-08), *Bacillus licheniformis* (BPT-20), *Bacillus licheniformis* (BPT-23) and *Bacillus pumilus* (BPT-25) were subjected to dual susbtrate plate assay for protease activity.

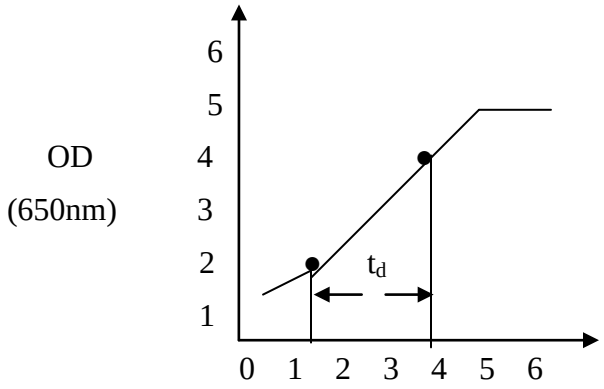
2.1.2.2.Growth studies of isolated from NIBGE culture collection

As these isolates had the origin from salt mines (Nasrin et al., 2008).. These bacteria were tested to grow on halophilic and alkaliphlic media to observe their extremophilic nature.

a. Alkaliphilic Medium (AP) Containing g/l: Yeast Extract 5, Polypeptone 5, KH_2PO_4 1, $MgSO_4\ 7H_2O$ 0.2, Na_2CO_3 10.

b. Halophilic Medium (HP) Containing g/l: Tryptone 10, Beef Extract 10, NaCl 10 or 65, and pH 7.4.

The growth pattern of all the isolates (BPT 05,06,08,20,23 & 25) were studied on both media respectively, by incubating at 37°C in 250 ml Erlenmeyer flask. For calculating the doubling time (t_d) and specific growth rate (μ) sample aliquots were taken from growing cultures at regular time intervals to record absorbance at 650nm in a double beam spectrophotometer. Then graphs were plotted by taking absorbance at Y-axis and time at X-axis. The doubling time was calculated for each culture as depicted in fig 2.1 and specific growth rate was calculated by using the following expression: $t_d = 0.693/\mu$



Time (hours)

Fig. 2.1: Graph showing measurement of doubling time.

2.1.2.3. Genomic DNA extraction from bacterial samples

The same protocol was used to isolate genomic DNA from the harvested microbial biomass as described in section no. 2.1.1.4.

2.1.3. Collection of metagenomic samples and their characterization

Three soil samples were taken in sterile falcon tubes from various sites (Table 2.1) to carry out metagenomic studies. Soil samples were made into slurries by mixing with sterile distilled water (50%, w/v) in orbital shaker (Kuhner Shaker) at 100 rpm for 24 hours at room temperature 30°C. Some physical characteristics of the samples were studied viz., EC, pH and temperature.

Table 2.1 Description of samples collected from various sites.

Sr. No	*Sampling site*	*Nature of the sample*	*Sample code*
1	NIBGE Canal bank	Muddy Soil	S_A
2	NIBGE cotton feild	Dry Soil	S_B
3	NIBGE Coal Heap	Dry Soil	S_C

1. Analysis of Na^+ and K^+

For Na^+ and K^+ analysis samples were first filtered using Whattman No. 3 filter papers to remove particulate material. Na^+ and K^+ levels (in mg/kg) were determined by flame photometer (Fp20 SEAC Radium group).

2. Analysis of other metals

For analysis of other metals like Copper, Iron, Zinc, Cadmium, Nickel, Magnesium, Manganese and Silver, 2 grams of respective soil sample were added to aquaregia (HCl $+H_2SO_4$), the solution was steam distilled until 10ml was left , washed it with distilled water and then filtered through Whattman filter paper, added the filtrate to a 200ml volumetric flask and made the volume upto the mark. Metal analysis (in mg/kg) was carried out by Fast Sequential Atomic Absorption spectrophotometer (AA240FS).

2.1.3.1. Protease activity of isolates metagenomic samples

All the three metagenomic samples were tested for protease activity by dual substrate plate assay , from the slurries of sample S_A, S_B and S_C as described in section no. 2.1.1.5.

2.2. MOLECULAR BIOLOGY STUDIES FOR PROTEASE GENE

Template DNA was isolated both from cultured organism and from metagenomics samples. The protocol used for cultured organism has been described in section no. 2.1.1.4. However, metagenomic samples required significantly different treatment for DNA acquisition.

2.2.1. Direct DNA extraction from soil samples

1. 100ml of extraction buffer containing 5mg proteinase K was mixed with 50g of each soil sample in 250 ml centrifuge tube.
2. The samples were incubated at 37°C for 30 min with shaking at 180 rpm.
3. Then 10 ml of SDS (20%) was added and the samples were incubated at 65°C for 90 minutes.
4. The supernatant was collected after centrifugation at 10000rpm for 10 minutes at room temperature.

5. Then the supernatant was transferred to 50 ml capacity centrifuge tubes for each soil sample, containing half volume of polyethylene glycol 30%/sodium chloride (1.6 M) and incubated at room temperature for 2 hours.
6. Samples were centrifuged at 12000 rpm for 20 minutes and partially purified nucleic acid pellets were resuspended in 10 ml of TE Buffer.
7. Potassium acetate (7.5m) was added to a final concentration of 0.5M
8. Samples were transferred to ice for 5 minutes then centrifuged at 1200rpm for 30 minutes at 4°C to precipitate protein and polysaccharides
9. The aqueous phase was extracted with phenol chloroform and chloroform isoamyl alcohol and DNA was precipitated by adding 0.6 volume isopropanol
10. After 2 hrs at room temperature DNA was pelleted by centrifuged at 14000rpm for 30 min and resuspended in 1ml TE buffer.
11. Quality of DNA was checked by running on 1% agarose gel. (as described above).

The next part of our studies was to carry out molecular biology studies on DNA obtained from bacterial isolates from NIBGE culture collection as well as DNA from metagenomic soil samples.

2.2.2. Primer designing and synthesis

To recover protease gene from genomic DNA by PCR, five pairs of primers were used (Table 2.1). Out of which three primers were selected from sequence of alkaline protease gene of Bacillus isolates already published in NCBI. In case of designing of primers we used different databases available on internet (www.expasy.ch and www.ncbi.nlm.nih.gov). New Primer sets were named as F1 & R1, F2 & R2, according to the nature of forward and reverse. On the other hand three already reported primers

(A1 & A2, B1 & B2, C1 & C2) were also used (Jiao *et al*., 2004, Tang *et al*., 2004). All the primers were synthesized commercially.

Primers were designed by taking into account following considerations are necessary to get specific amplifications with high yields. A brief description is given below.

1. Primer Length: It is generally observed that optimal length of PCR primer is 18-22 bp, this length is long enough for adequate specificity and short enough to bind easily to the template at annealing temperature.
2. Primer Melting Temperature (Tm): By definition it is the temperature at which one half of the DNA duplex will dissociate to become single stranded and indicates duplex stability. Primers with Tm in range of 52- 58^{o}C generally produce best results. Those with Tm above 65^{o}C have a tendency for secondary annealing. The GC content of the sequence give fair indication of primer Tm.
3. Primer Anneling Temperature: The primer Tm is the estimate of the DNA-DNA hybrid stability and critical in determining the annealing temperature. Too high Ta will produce insufficient primer-template hybridization resulting in low PCR product yield. Too low Ta may possible leat to non-specific products caused by a high number of base pair mismatches. Mismatch tolerance if found have strong influence on PCR specificity.
4. GC Content: The GC content (the number of G's and C'c in the primer as a percentage of the total bases) of primer should be 40-60%.
5. GC Clamp: The presence of G or C bases within five bases from the 3' end of primers (GC clamp) helps promote specific binding at the 3'due to the stronger bonding of G and C bases. More than 3 G's or C's should be avoided in the last 5 bases at the 3' end of the primer.

6. Primer Design Using Softwares: A number of primer design tools are available that can assist in PCR primer design. These tools may reduce the time and cost involved in experimentation also by lowering chances of failed experimantion.

2.2.3. PCR amplification of protease gene

Isolated DNA from bacterial isolates and metagenomic samples were subjected to PCR amplification using primers given in Table 2.2:

Table 2.2 : Nucleotide sequences of primer sets used

Primers	Sequence	Purpose
A1 A2	5-TTTCCAAGCGACTTAATTCC-3 5-CGCGTCGACGGCATCAAGAACCGTGCAGC-3	For intact AP gene
B1 B2	5-AGTCCATGGCCCAAACCGTCCCTTAT-3 5-TCGGATCCTTAGTTAGAAGCTGCTTGAAC-3	For mature protease coding region
C1 C2	5-CATGCCATGGATGGTGAGGAAGAAGATTTTTGGC-3 5-GAATCTCGAGTTGAGCGGCAGCTTCGACATTG-3	Harbouring the start and stop codon region of the single peptide of B.licheniformis 6816 subtilisin protease gene
F1 R1	5-CGACCCAGCAAACAAGAGAGC-3 5-CCTCTGCCGTAAGCATTAGCC-3	For complete gene sequence of alkaline protease
F2 R2	5-CCGGGTTTCCCAATAGGC-3 5-CGTGTGCCCTTTAACGCAC-3	For complete gene sequence of alkaline protease

Reaction mixtures consisted of approximately 5 ng of template DNA, 100 pmol of each of the primer, 200 nM deoxynucleoside triphosphate (dNTP) mix, 2 mM $MgCl_2$ and 1U of Taq DNA polymerase.

For 25µl Reaction Mixture

1) 13 µl of deionized distilled water was taken in 0.2 ml Eppendorf.
2) Added 2.5 µl of 1X PCR buffer
3) 2 µl of $MgCl_2$
4) 0.2 µl of dNTPs
5) Then added 1 µl of forward and reverse primers each

6) 5 µl template DNA

7) 0.3 µl of Taq polymerase was added at the end.

2.2.4. PCR profile for protease gene amplification

Amplification of protease gene was carried out under following conditions: 94^{o}C for 5 minutes for 1 cycle, 94^{o}C for 1 minute, 56^{o}C for 1 minute, 72^{o}C for 2 minutes for 32 cycles, and 72 ^{o}C for 10 minutes for 1 cycle.

2.2.5. Gel extraction

Gel extraction was done by using QIAquick® Gel Extraction Kit Instruction Manual in following steps.

1) Excised the DNA fragments from the agarose gel with a clean, sharp scalper.
2) Weighed the gel in a colourless tube. Added 3 volume of buffer QG to 1 volume of gel (100 mg ~100 µl).
3) Incubated at 50^{0}C for 10 minutes (or until the gel has completely dissolved). Gel can be dissolved by vortexing the tube every 2-3 minutes during the incubation.
4) After the gel dissolved completely, checked the colour of the mixture. It should be yellow. (similar to buffer QG without dissolved agarose).
5) Added 1 gel volume of isopropanol to the sample and mixture (This step increased the yield of DNA fragment <500bp and >4 kb only).
6) Placed a spin column in a 2 ml collection tube.
7) To bind DNA, applied the sample to the QIAquick column, and centrifuged for 1 minute.
8) Discarded the flow-through and placed QIAquick column in the same collection tube.
9) To wash, added 0.75 ml of buffer PE to QIAquick column and centrifuge for 1 minute.

10) Discarded the flow-through and centrifuged the QIAquick column for an additional 1 minute at 13,000 rpm.
11) Placed the QIAquick column into a clean 1.5 ml micro centrifuge tube.
12) To elute DNA, added 50 µl of buffer EB (10 mM Tris.cl,pH8.5) or H_2O to the centre of the QIAquick membrane and centrifuged the column for 1 minute. Alternatively, for increased DNA concentration, added 30 µl elution buffer to the centre of the QIAquick membrane, let the column stand for 1 minute, and then centrifuged for 1 minute.

2.2.6. Cloning of protease gene

2.2.6.1.Preparation of LB plates with Ampicillin

The medium was prepared by adding 15g agar to 1 liter of LB medium. Which was then autoclaved and allowed the medium to cool to 50°C before adding ampicillin to a final concentration of 100µg/ml. 30–35ml of medium was poured into 85mm Petri dishes and let the agar harden. Which can be stored at 4°C for up to 1 month or at room temperature for up to 1 week.

2.2.6.2.Preparation of LB plates with Ampicillin/IPTG/X-Gal

LB plates were made with ampicillin as above; then supplemented with 0.5 mM IPTG and 80 µg/ml X-Gal and plates were poured. Alternatively, 100 µl of 100 mM IPTG and 20 µl of 50 mg/ml X-Gal might be spreaded over the surface of an LB/ampicillin plate and allowed to absorb for 30 minutes at 37°C prior to use.

2.2.6.3.Ligation using the InsT/Aclone™ PCR product cloning kit

Plasmid vector pTZ57R/T DNA	3ul
PCR fragment	5ul
10X ligation buffer	3ul

PEG 4000 solution	3ul
T4 DNA ligase	1ul
Deionized distilled water	5ul

This reaction mixture was incubated at 16 ^{o}C for four hours. For maximum yield of useful recombinants, the reaction time can be extended to overnight.

2.2.6.4.Preparation of heat shock competent cells of *E.coli.* (10b strain)

Following protocol was used for the competent cells formation:

1) A single colony from a freshly grown plate of *E.coli* was picked and transformed into 50 ml LB medium in 250 ml flask and incubated at 37 ^{o}C overnight with vigorous shaking.
2) 2.5 ml of the overnight culture was taken and diluted to 250 ml in 1 liter flask and shaken vigorously at 37 ^{o}C until OD of 0.5-1.0 (10^{10} cells/ml).
3) Culture was cooled by placing on ice for 30 minutes. The cells are transferred aseptically to sterile disposable 50 ml propylene tubes.
4) The cells were pelleted by centrifugation at 5000 rpm at 4 ^{o}C for five minutes and resuspended in 15 ml of 0.1M $MgCl_{2.}$
5) The cells were pelleted by centrifugation at 5000 rpm at 4 ^{o}C for 5 minutes and resuspended in 15 ml of 0.1M $CaCl_2$ and kept on ice for 30 minutes.
6) The cells were again pelleted by centrifugation at 5000 rpm for 5 minutes and suspended finally in appropriate amount of 0.1M $CaCl_2$ and sterile cold 100% glycerol.
7) The cells were stored in aliquots of 200 ul at –70 ^{o}C.

2.2.6.5. Transformation using heat shock competent cells of *E.coli*

Following protocol was used for the transformation of PCR product using heat shock competent cells:

1) 5-ul plasmid was added in 200-µl cells aliquot and kept on ice for half an hour.
2) Heat shock was given on dry bath at 42 ^{o}C for 2 minutes.
3) Cells were kept on ice for 2 minutes.
4) 1 ml LB was added in the mixture.
5) Cells were kept at 37 ^{o}C for one hour.
6) Cells were pelleted by centrifugation at 13000 rpm for 20 seconds and supernatant was removed.
7) Dissolved the pellet in 200 ul remaining supernatant and spreaded on the selection LB/ampicilline/IPTG/X-GAL agar plates.
8) Kept the plates overnight at 37 ^{o}C.
9) Growth appeared in the form of blue and white colored colonies. LB liquid media were prepared and autoclaved and after cooling added 5µl of ampicillin in concentration of 100 mg/ml in it and poured 5 ml of medium in each test tube. Picked up single white colony from cultured plates with the help of sterilized pipette tips and inoculated the liquid medium. Test tubes were placed overnight in shaker at 37 ^{o}C.

2.2.6.6.Plasmid isolation

Plasmid isolation was carried out by using Quantum prep® Plasmid Miniprep Kit Instruction Manual.

1) All centrifugation steps were performed at maximum speed (12-14,000 rpm)
2) Transfered an overnight culture (1-2 ml) of plasmid containing cells to a micro centrifuge tube. Pellet the cells by centrifugation for 30 seconds. Removed all the supernatant by aspirating or pipetting.

3) Added 200 µl of the cell resuspension solution and vortex or pipeted up and down until the cell pellet was completely resuspended.
4) Added 250 µl of cell lysis Solution and mixed by gently inverting the caped tubes about 10 times (do not vortex). The solution should become viscous and slightly clear if cell lysis has occurred.
5) Added 250 µl of neutralization solution and mixed by gently inverting the caped tubes about 10 times (do not vortex).
6) Pelleted the debris for 5 minutes in a microcentrifuge. A compact white debris pellet formed along the side or bottom of tube. The supernatant at this step contained the plasmid DNA.
7) Transfered the clear lysate to a spin filter, added 200 µl of thoroughly suspended matrix, and then pipetted up and down to mix. If you have multiple samples, transfer the lysate first, then add matrix. When matrix has been added to samples and mixed, centrifuge for 30 seconds.
8) Removed the spin filter from tube, discarded the filtrate at the bottom of tube and replaced the filter in the same tube. Added 500 µl of wash buffer and washed the matrix by centrifugation for 30 seconds.
9) Removed the spin filter from tube, discarded the filtrate at the bottom of the tube and replace the filter in the same tube. Added 500 µl of wash buffer and washed the matrix by centrifugation for 2 minutes to remove residual traces of ethanol.
10) Removed the spin filter and discarded the micro centrifuge tube. Placed the filter in one of the 1.5 ml collection tubes, which accommodated the spin filter. Added 100 µl of deionized water or TE. Eluted the DNA by centrifugation for 1minute at top speed.

11) Discarded spin filter and stored the eluted DNA at –20^{o}C.

2.2.6.7. Digestion of miniprep

Digestion of miniprep was carried out as following:

Reaction mixture contained

Reaction buffer	2.0 µl
Plasmid DNA	10 µl
EcoR1	0.5 µl
HindII1	0.5 µl
Nanopure water	7.0 µl
Total volume	20 µl

Incubated the reaction mixture at 37°C for 2-2.5 hours.

2.2.6.8.Gel electrophoresis

Loading buffer (3µl) was mixed with reaction mixture and ran on 1.0% agarose gel at 70 Volts.

2.2.6.9.Selection of clones containing inserts/ PCR fragments for sequencing

Selection of clones was carried out after digesting plasmid DNA with *EcoR1* and *HindIII* restriction enzymes. Only clones, which contained an observable insert after the digestion reaction, were sequenced. DNA samples were sequenced with M13 forward primer. Those PCR products which could not be cloned were subjected to sequencing reaction as such using their respective protease gene amplification primers.

2.2.7. Protease gene sequencing

Cloned and amplified PCR products of protease gene both from bacterial stains and metagenomic sample were used for sequencing in Beckman Coulter Sequencer at NIBGE. To confirm the validity of our results these samples were also sequenced commercially.

Dye Terminator cycle sequencing with Quick start kit

1. **Prepearation of the DNA sequencing reaction**

dH_2O (to adjust volume to 20µl	0-9.5 µl
DNA template	0.5-10 µl
Primer	2.0 µl
DTCs Quick start master Mix	8.0 µl
Total	20 µl

2. **Thermal cycling programme**

96^{o}C	20 sec
50^{o}C	20 sec
60^{o}C	4min

For 30 cycles followed by holding at 4^{o}C

3. **Ethanol precipitation**

Precipitation in individual tubes:

a) Prepare a labeled, sterile 0.5 µl microfuge tube for each sample.

b) To each labeled tube. Add fresh 4 µl stop solution (equal volume of 3M NaOAc pH 5.2 + 100mM EDTA pH 8.0) and 1 µl of 20mg/ml glycogen (supplied with kit)

c) Transfer the sequencing reaction to the appropriately labeled 0.5ml microfuge and mix thouroghly.

d) Add 60 µl cold 95% (v/v) ethanol/ dH_2O from -20^oC freezer and mix thoroughly.

e) Carefully remove the supernatant with a micropipette (the pellet should be visible).

f) Rinse the pellet 2 times with 200µl 70% (v/v) ethanol/ dH_2O from -20^oC freezer. For each rinse, centrifuge immediately at 14,000 rpm at 4 ^oC for a minimum of 2 mins. After centrifugation carefully remove all of the supernantant with a micropipette.

g) Vaccum dry for 10 minutes (or until dry)

h) Resuspended the sample in 40µl of the sample loading solution (SLS) provided in kit.

4. Sample preparation for loading into instrument

a. Transfer the resuspended samples to the appropriate wells of the sample plate.

b. Overlay each of the resuspended samples with one drop of light mineral oil.

c. Load the sample plate into the instrument.

DTCS quick start master mix (composition)

10X sequencing buffer	200µl
dNTP mix	100µl
ddUTP Dye Terminator	200µl
ddGTP Dye Termintor	200µl
ddCTP Dye Terminator	200µl
ddATP Dye Terminator	200µl
DNA polymerase	100µl

2.3. BIOINFORMATICS STUDIES ON PROTEASE GENE

Protease gene sequences obtained from both bacterial isolates and metagenomic sample were analyzed using different bioinformatics tools available on internet,using both DNA and Protein databases also various softwares.

2.3.1. DNA sequence analysis

2.3.1.1.Sequence homology search

Protease gene sequence homology searches were carried out databases using BLAST algorithm (McGinnis *et al*., 2004) at the National Center for Biotechnology Information (USA) against sequences entries in GenBank (www.ncbi.nlm.nih.gov). Homologues of nucleotide sequences were found using BLAST N and were used to build pair wise sequence alignment between the target and template gene sequences.

2.3.1.2.Multiple sequence alignment

Comparison of our protease gene sequences with protease genes sequences from selected organisms was carried out by obtaining sequences from GenBank. For each of our protease gene sequence multiple sequences were obtained and were aligned using CLUSTALX program(version 1.81) (Thompson *et al*., 1997)with default parameters and manually adjusted where necessary. Multiple alignment between our protease sequences was also carried out to check the sequence similarity between them. It helps to identify functionally important regions, phylogenetic reconstruction and build sequence profile.

2.3.1.3.Phylogenetic analysis

The aligned sequences were used to construct a distance matrix (Jukes and Cantor 1969), after the generation of 100 bootstrap sets, that was subsequently used to construct

a phylogenetic tree by neighbor-joining method (Saitou and Nei 1987), using PHYLIP (Felsenstein,J ,1992)Program . Multiple sequence alignment file was used to construct the structural relationship among DNA sequences.

2.3.1.4.Open reading frame

ORF Finder programme at NCBI (http://www.ncbi.nlm.nih.gov) ORF cognitor compares ORFs with the clusters of orthologous groups (COG) protein database was used to identify the open reading frames in our sequences.

2.3.1.5.Restriction map/ GC content/ Nucleotide sequence statistics

To identify various restriction sites on our protease gene sequences restriction maps for all were constructed using CLCbio main work bench software. GC content and Nucleotide sequence statistics in our protease gene sequences were determined by using CLCbio DNA work bench software.

2.3.1.6.Translation of DNA to protein

DNA sequences were then subjected to Translate program of Expasy (Gasteiger *et al*.,2003) to get the translated Protein sequences. (http://www.us.expasy.org).

2.3.2. Protein sequence analysis

2.3.2.1.Protein pairwise alignment and template selection

Primary protein sequences of all the target proteins were subjected to BLAST (Psi-BLAST) algorithm (McGinnis *et al*., 2004) against Protein Data Bank (PDB) was used to carry out the sequence homology searches. The templates were chosen on the basis of highest sequence identities between the targets and templates. The crystal structure coordinates of 1AVT, 1BH6, 1C3L, 1MEE, 1SCJ has been selected as templates for

constructing the comparative models of these proteases. The 3D coordinates of templates were extracted from the Brookhaven Protein Data Bank (http://www.rcsb.org/pdb) (Berman *et al*., 2000).

2.3.2.2.Multiple sequence alignment

To determine protein family classification and phylogenetic tree reconstruction, multiple sequence alignment was carried out using the program CLUSTAL X (Version: 1.81) (Thompson *et al*., 1997) with default parameters and manually adjusted where necessary. Sequences homologous to the target sequences were extracted from SWISS-PROT ((Bairoch *et al* ., 2005).

2.3.2.3.Phylogenetic analysis

By the construction of multiple sequence alignment a matrix of pairwise similarity scores is calculated which can be employed to construct phylogenetic tree to express the structural relationship among the protein sequences in the family. Phylogenetic lineage was calculated using the phylogeny Inference program Phylip (Version 3.6b) (Felsenstein *et al.,* 1992).

2.3.2.4.Secondary structure prediction

The Secondary structure prediction of our protease sequences was carried out by submitting the sequence to Consensus Secondary Structure Prediction Sever at http://pbil.ibcp.fr/NPSA/npsa_npsa.html (Deléage *et al*., 1997). This gives us the prediction regarding the presence of alpha helices, beta sheets and loop regions present in our sequence.

2.3.2.5. Protein model building and refinement

Three dimensional homology models of these proteases were constructed using the crystal coordinates of templates on the basis of alignment between target and template sequences. All steps of homology modeling and refinement were carried out by using the SwissModel server at http://swissmodel.expasy.org (Arnold *et al.*, 2006).

2.3.2.6.Model Visualization

To examine the reliability of the alignment and modeling of variable surface loops, structural investigations on the graphics screen 3D visualization programs, Ds-Viewer (Pazal.D.P, 1989) and Astexviewer (Hartshorn, M.J, 2002) were used.

2.3.2.7.Model Evaluation

For the evaluation of the predicted model analysis of geometry, stereochemistry, and energy distributions in the models was carried out. Protein structure validation software suite (PSVS) is used for the assessment of protein structures generated from Homology modeling.which gives the reliability of the homology model by PROCHECK program (Laskowski *et al*., 1996) (Version: 3.4). The principle of PROCHECK is to evaluate the overall stereo-chemical quality of a given model. It checks the geometry of the residues in a given protein structure as compared with well-refined structures at the same resolution. The energetic architecture of protein folds was determined by using the web interface of ProSA i.e. ProSA-Web at https://prosa.services.came.sbg.ac.at/prosa.php (sipple *et al*., 2007). ProSA-web provides an easy-to-use interface to the program ProSA (Sippl *et al*., 2007) which is frequently employed in protein structure validation. ProSA calculates an overall quality score for a specific input structure. It gives an energy graph which provides diagnostic tools to mark problematic segments. Energy graphs also

demonstrate the energetic architecture of protein folds as a function of amino acid sequence position. It calculates a score for input-structure. Scores of native protein folds are in a distinctive range. If score is outside this range then the structure may have problems.

CHAPTER

3.0

RESULTS

3.1. MICROBIOLOGICAL STUDIES FOR PROTEASE GENE

Microbiological studies were carried out on samples obtained from three sources:

1. Environmental samples from pesticide factory
2. Bacterial isolates from NIBGE culture collection
3. Soil samples from NIBGE for metagenomic studies

3.1.1. Collection of environmental samples

Soil samples were collected in sterilized glass bottles from the sale point of insecticide as well as from different locations of a canal, receiving drainage water of a factory involved in the formulation of pesticides (Fig. 3.1.1).

Fig.3.1.1. Photograph **(arrow)** showing site of sample collection.

3.1.1.1. Isolation of microorganisms from environmental samples

Soil samples obtained from pesticide factory waste yielded small colonies over the surface of nutrient agar.(Fig.3.1.2).

3.1.1.2. Characterization and diversity studies of the isolated bacteria

Selection of bacterial strains was made on the basis of colony characteristics and bacterial morphology which indicated the microbial diversity of microorganism present in the respective environmental samples.

3.1.1.3. Colony characteristics of different isolates

On the basis of colony color, isolates were distinguished from one another. Colonial pigmentation of these samples included yellow, cream and white. These colonies were mostly irregular and circular in form. Colony diversity was evident when the colony elevation was considered and was found to be raised and flat. (Table 3.1.1). Most of the colonies had entire and undulate margins.

3.1.1.4. Microscopic observations

Different isolates displayed different cellular morphologies when viewed under the phase contrast microscope. They were mostly rod shaped except isolate S2. They were indicated to contain no endospores. None of the isolates indicated exospores. All isolates were motile in nature except isolate S2 and all were Gram's negative. These results have been tabulated in Table 3.1.1

Table 3.1.1. Characteristics of isolates from soil of pesticide factory waste.

Isolate code	**Media used**	**Shape of cells**	**Cell size (µm)**	**Motility**	**Gram's staining**	**Endo-spores**	**Colony shape**	**Colony margins**
2A	Nutrient Agar	rods	1.5-3.0	Motile	Negative	Negative	Circular	Entire
3A	-do-	rods	1.5	Motile	-do-	-do-	Irregular	Entire
3B	-do-	rods	2.0	Motile	-do-	-do-	Irregular	Undulate
S1	-do-	rods	2.5	Motile	-do-	-do-	Irregular	Undulate
S2	-do-	Coccus/rod	1.5	Non motile	-do-	-do-	Circular	Undulate

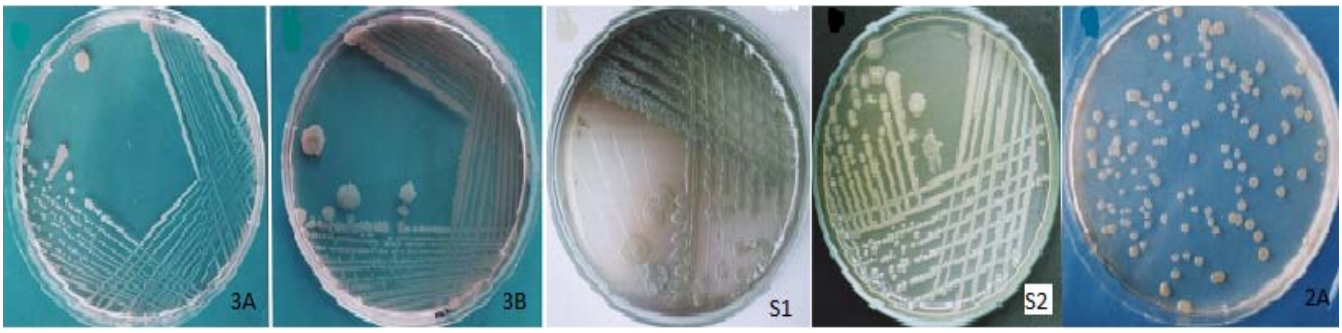


Fig. 3.1.2. Appearance of colonies of the bacterial isolates **3A**, **3B, S1, S2** and **2A** on nutrient agar.

3.1.1.5. Molecular characterization of the isolates from environmental samples

3.1.1.5.1. DNA isolation

Respective colonies were picked and were enriched to obtain appropriate biomass. These pure cultures were then used for DNA isolation. Isolated DNA was visualized on 1% agarose gel (Fig. 3.1.3).

Fig: 3.1.3. DNA isolated from Environmental samples.
Lanes: 1, 2A; 2, 3A; 3, 3B; 4,S1; 5,S2.

3.1.1.5.2. PCR amplification of the 16S rDNA and nucleotide sequencing

DNA samples were amplified after optimizing the PCR conditions. Using FD1 and rp1 primers a PCR product of 1,500 bp was obtained in all the cases (Fig. 3.1.4). Purified PCR products were partially sequenced commercially and the sequences that were obtained were very

fine and only few Ns were present in some sequences (Appendix I). According to nucleotide BLASTN (NCBI, http://www.ncbi.nlm.nih.gov) results for the partial 16S rDNA sequence showed that isolates 2A,3A,3B,S1 had homology with *Pseudomonas* spp. And isolate S2, which had homology with *Acinetobacter baumannii*. Sequence identity values were 99 % for 2A, 97% for 3A ,98% for 3B and S1,97% and gap percentage was zero for 2A,3A and 3B,6% for S1 and 1% for S2 sequence (Table 3.1.2).

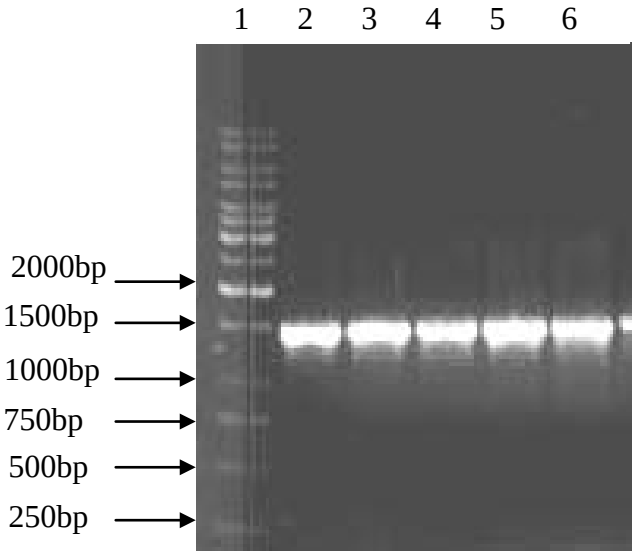


Fig.3.1.4. PCR amplification of 16S rDNA of different isolates
Lanes:1, 1Kb, DNA marker; 2, 2A; 3, 3A; 4, 3B; 5, S1; 6, S2

Table 3.1.2. Sequence identity (%) of the partial 16S rDNA gene sequence of isolates obtained from environmental samples with related microbes.

Isolatewith Accession No.	Base pairs	Top Hit with	% Identities	Gap %
2A	823	*Pseudomonas aeruginosa* (AB449026)	99	0
3A	810	*Pseudomonas sp. HF3-3* (DQ288108)	97	0
3B	818	*Pseudomonas sp. DQ-02* (GU325691)	98	0
S1	809	*Pseudomonas sp.* P1 (GU113071)	98	6
S2	808	*Acinetobacter baumannii AB0057* (CP001182)	97	1

3.1.1.5.3. Phylogenetic analysis

In order to determine the phylogenetic position of the isolates, the partial 16S rDNA gene sequences were compared with strains of *Pseudomonas* species retreived from GenBank at NCBI (www.ncbi.nlm.nih.gov). (Fig 3.1.5).

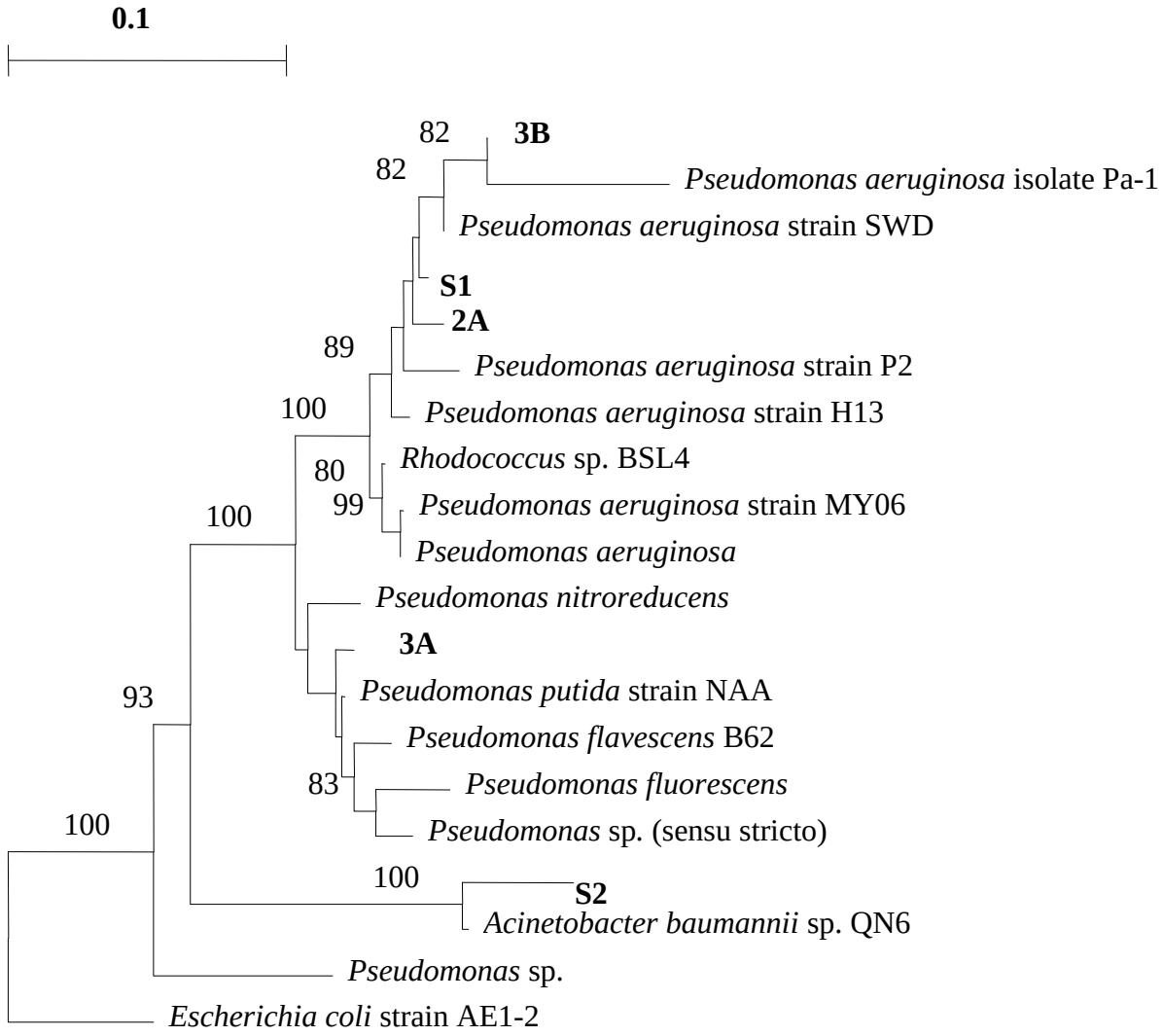


Fig.3.1.5. The inferred relationship, based on partial 16S rDNA sequences, isolated in this study (in bold) to other bacteria. The tree was rooted with *E.coli*. Scale bar represents the number of inferred nucleotide substitution per site. Bootstrap values (100 replicates) are shown at the nodes.

3.1.1.6. Protease activity assay for cultured organism environmental samples

The organism identified through 16S rDNA analysis were tested for protease activity using a bioassay developed in NIBGE laboratories but no significant results were found. Therefore further studies were not conducted on DNA of these organisms.

3.1.2. Bacterial isolates from NIBGE culture collection

3.1.2.1. Protease activity of isolates from NIBGE culture collection

All the six pure extremophilic bacterial isolates i.,e, *Bacillus amyloliquefaciens* (BPT-05), *Bacillus cereus* (BPT-06), *Staphylococcus arlettae* (BPT-08), *Bacillus licheniformis* (BPT-20), *Bacillus licheniformis* (BPT-23) and *Bacillus pumilus* (BPT-25) were subjected to dual substrate plate assay for protease activity. All of these isolates were found protease positive, which was indicated by the formation of a clear zone around the spot inoculation (Fig 3.1.6). These isolates were further used for the amplification and characterization of protease gene.

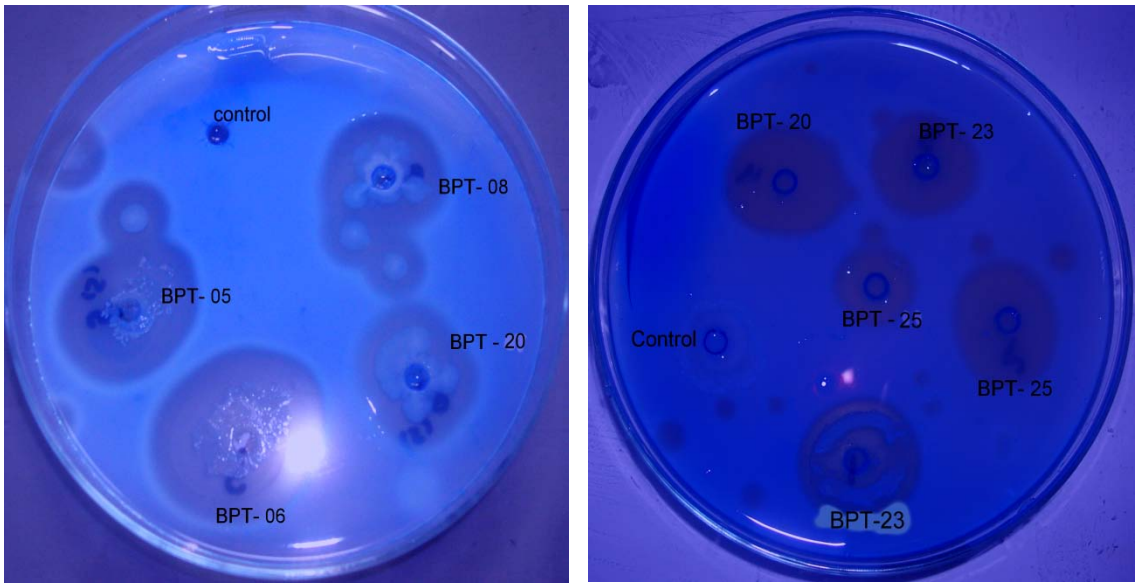


Fig.3.1.6. Protease activity assay on bacterial samples along with control.

3.1.2.2. Colony characteristics and morphology of bacteria

Isolates obtained from NIBGE culture collection were originally isolated by Nasrin *et al* (2008), According to these studies all the isolates displayed different cell morphologies when viewed under the phase contrast microscope.

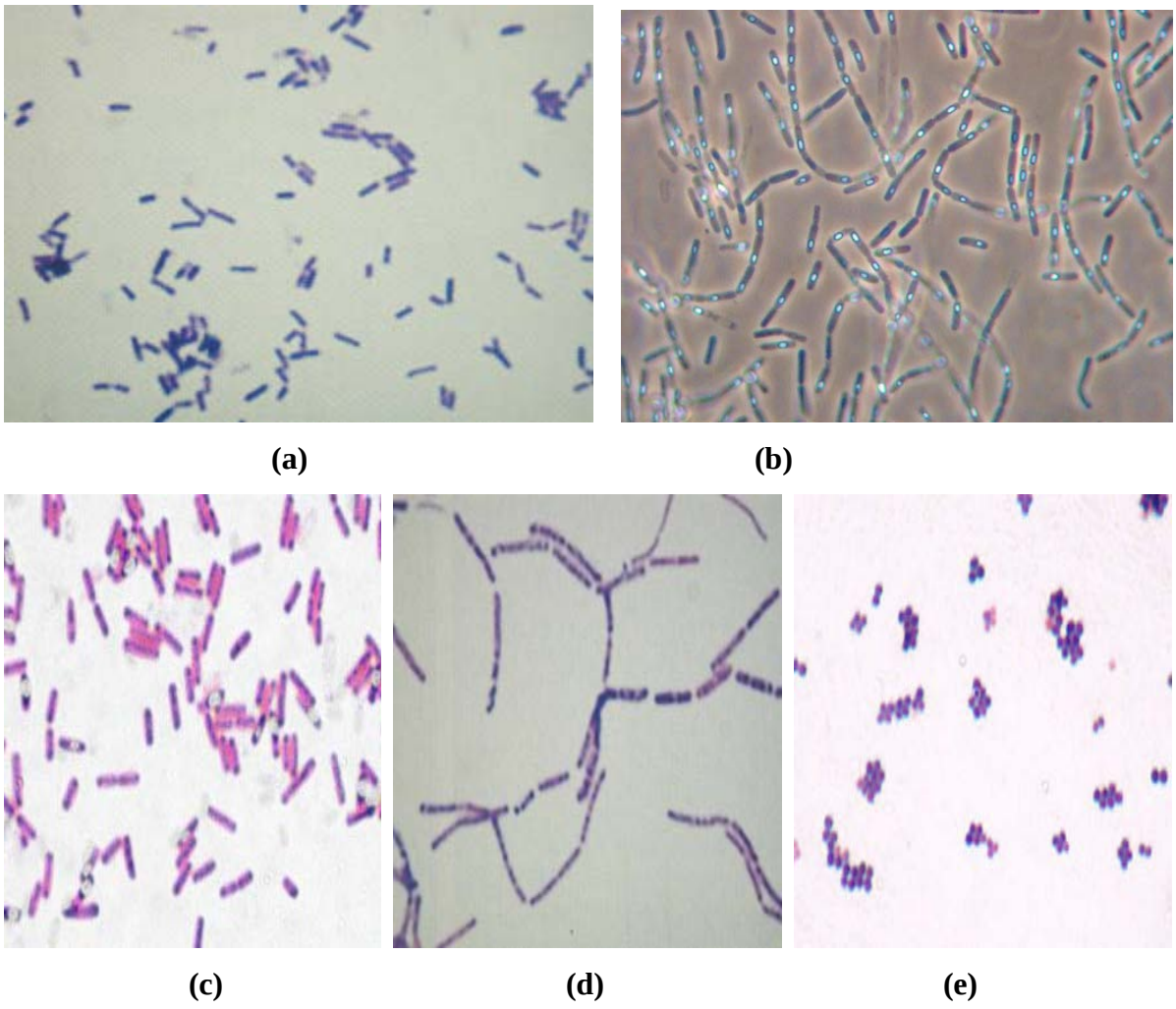


Fig.3.1.7. Phase contrast microscopic Images of bacteria isolates (a) *Bacillus amyloliquefaciens*, (b) *Bacillus cereus*, (c) *Bacillus licheniformis*, (d) *Bacillus pumilus*, (e)*Staphylococcus arlette.*

These cells were mostly rod shape except BPT-08. Three of the isolates BPT-08, 23, 25 had circular colony shape and the other three BPT-05, 06, 20 had irregular colony shape with raised

colony elevations. BPT-06 was indicated to contain endospores.Table.3.1.3 showed the detailed characteristic of these isolates (Nasrin *et al*., 2008).

Table.3.1.3 Characterictics of the isolates obtained from NIBGE culture collection.
(Nasrin *et al*., 2008).

Isolate Code	**Colony Color**	**Colony Form**	**Colony Elevation**	**Colony Margins**	**Bacterial Shape and Size (μm)**
BPT-5	White	Irregular	Flat	Undulate	Rods, 1-1.5
BPT-6	Red	Irregular	Umbonate	Undulate	Rods, 1-1.5
BPT-8	White	Circular	Raised	Entire	Coccus, 0.5-1
BPT-20	Yellow	Irregular	Raised	Lobate	Rods, 2-2.5
BPT-23	Yellow	Circular	Raised	Entire	Rods, 2-3
BPT-25	White	Circular	Raised	Entire	Rods, 1

3.1.2.3.Growth studies in halophilic media

Growth studies of all the isolates were carried using halophilic media. All the isolates displayed different behavior from one another. All of these isolates had a prolonged lag phase of 3-5 hours at 3% NaCl concentration and 10-17 hours in 6% NaCl containing media. Almost all of the isolates displayed non significant growth at 9% salt concentration except from BPT-23.

Table 3.1.4. Doubling time and specific growth rate for the isolates taken from NIBGE culture collection at various salt concentrations in halophilic media.

Isolate code*	Doubling time td (hr)			Specific Growth rate μ (h^{-1})		
	3% NaCl	6%NaCl	9%NaCl	3% NaCl	6%NaCl	9%NaCl
BPT-05	1.5	2	No growth	0.57	0.34	-----
BPT-06	1.15	2	No growth	0.69	0.34	-----
BPT-08	1.15	2	No growth	0.55	0.34	-----
BPT-20	2	2.7	No growth	0.34	0.27	-----
BPT-23	2.5	1.5	0.65	0.27	0.55	0.59
BPT-25	3.25	2.5	1.5	0.21	0.27	0.55

*Full names- *Bacillus amyloliquefaciens* (BPT- 05), *Bacillus cereus* (BPT-06), *Staphylococcus arlettae* (BPT-08), *Bacillus licheniformis* (BPT-20), *Bacillus licheniformis* (BPT-23) and *Bacillus pumilus* (BPT-25)

a) Isolate BPT-05

Bacillus amyloliquefaciens show maximum growth after 19 hours of incubation at 37°C in 3% and 6 % NaCl concentration in the liquid medium. It did not showed significant growth in 9% NaCl concentration. The growth curve was drawn for measuring the doubling time with the help of various softwares (Table. 3.1.4). These results showed that it can be placed in the catogery of modrate halophiles (Fig.3.1.8).

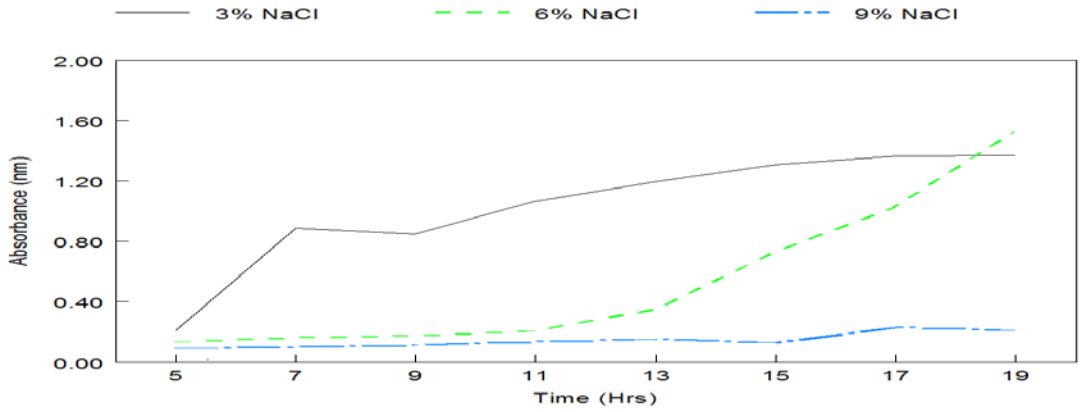


Fig.3.1.8. Growth curve of BPT-05 in halophilic media at different salt concentrations.

b) Isolate BPT-06

Bacillus cereus also showed maximum growth after 19 hours of incubation at 37°C in 3% and 6 % NaCl concentration in the liquid medium. Like BPT-05 it did not showed significant growth in 9% salt media. (Fig.3.1.9). The results showed modrate halophilic character of this strain. Its doubling time and specific growth rate are shown in Table 3.1.4.

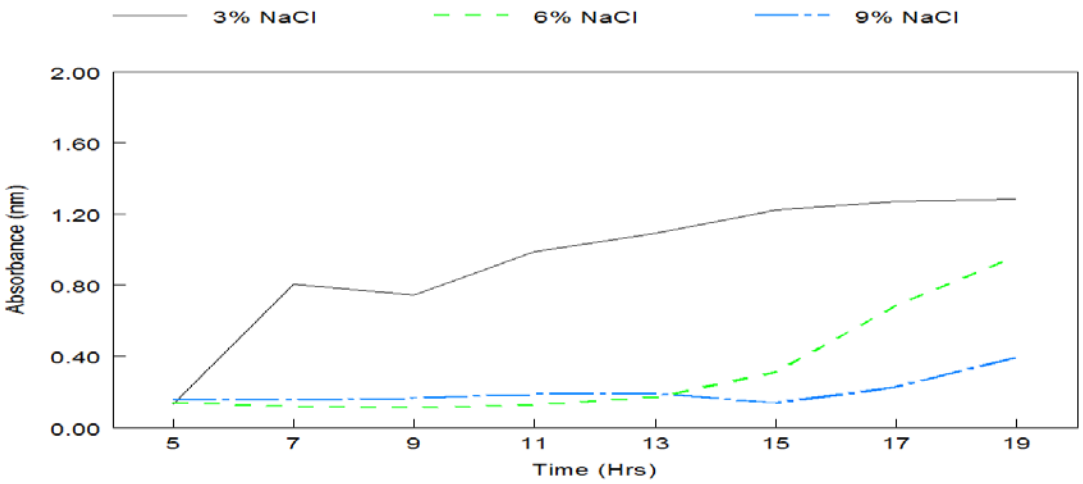


Fig.3.1.9. Growth curve of BPT-06 in halophilic media at different salt concentrations.

c) Isolate BPT-08

Staphylococcus arlettae indicated maximum growth after 19 and 21 hours of incubation at 37°C in 3% and 6 % NaCl concentration respectively. It did not show considerable growth in 9% salt media (Fig.3.1.10). This strain can also be categorized as modrate halophile.Doubling time and specific growth rate of this bacteria at various salt concentration are given in Table 3.1.4.

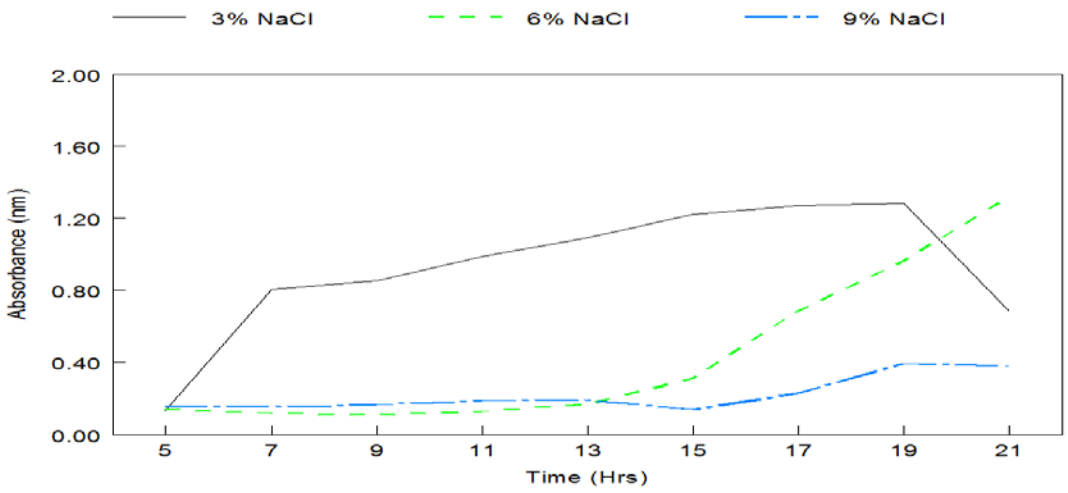


Fig.3.1.10. Growth curve of BPT-08 in halophilic media at different salt concentrations.

d) Isolate BPT-20

Bacillus licheniformis grew well after 19 and 21 hours of incubation at 37 °C in 3% and 6 % NaCl concentration respectively. It also did not showed considerable growth in 9% salt media (Fig.3.11) showing moderately halophilic character. Doubling time and specific growth rate of this bacteria at various salt concentration are given in Table 3.4.

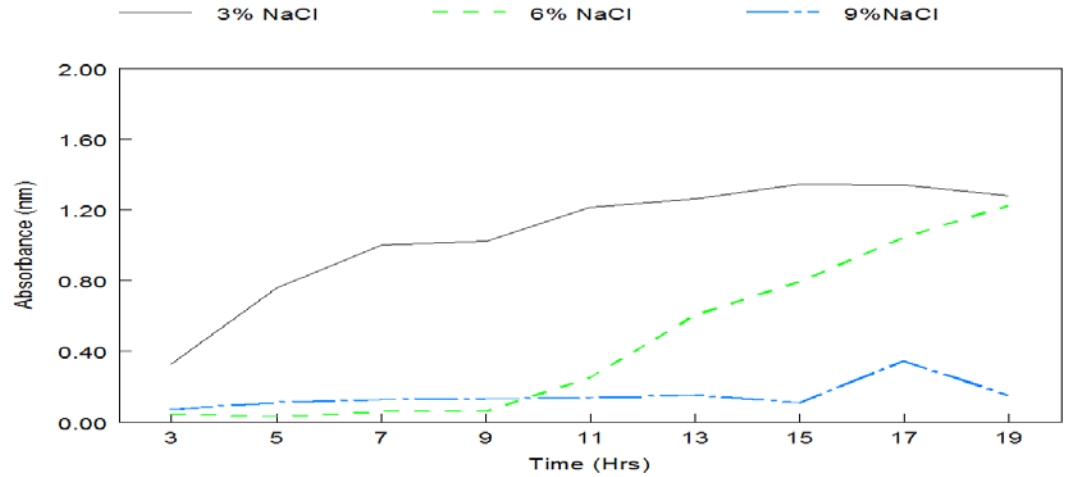


Fig.3.1.11. Growth curve of BPT-20 in halophilic media at different salt concentrations.

e) Isolate BPT-23

Bacillus licheniformis presented maximum growth after 19 and 21 hours of incubation at 37 °C in 3% ,6% and 9% NaCl concentration respectively (Fig.3.12). This isolate showed

extreme halophilic nature. Its doubling time and specific growth rate of this bacteria at various salt concentration are given in Table 3.4.

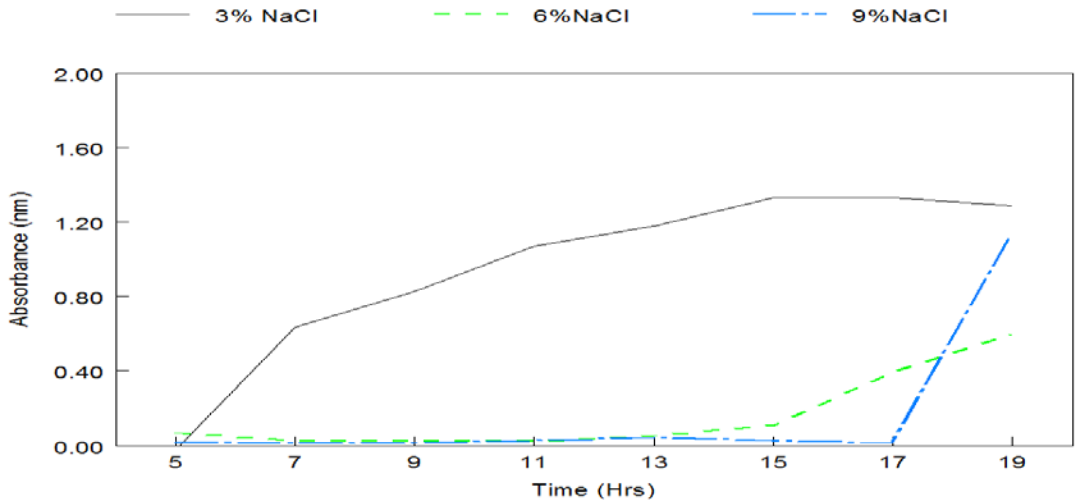


Fig.3.1.12. Growth curve of BPT-23 in halophilic media at different salt concentrations.

f) Isolate BPT-25

Bacillus Pumilus showed maximum growth after 19 and 21 hours of incubation in 3% and 6 % NaCl concentration respectively. It did not show considerable growth in 9% salt media (Fig.3.13). Doubling time and specific growth rate of this bacteria at various salt concentration are given in Table 3.4.

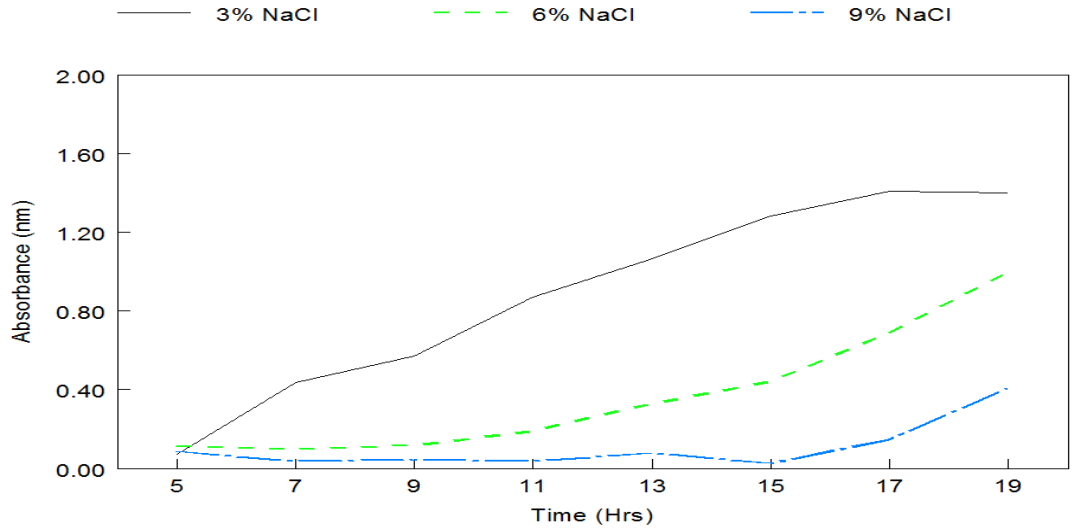


Fig.3.1.13. Growth curve of BPT-25 in halophilic media at different salt concentrations.

3.1.2.4. Growth studies in alkaliphilic media

Growth behavior on alkaliphilic media was observed for all the isolates, all of them had variable activity. The doubling times and specific growth rate for all isolates were calculated using sigma plot graphic program (Table.3.1.5).

Table 3.1.5. Specific growth rate and doubling time for the isolates taken from NIBGE culture collection at alkaline pH values(9,10 and 11).

Isolate code*	Doubling time td (hr)			Specific Growth rate μ (h^{-1})		
	pH9	pH10	pH11	pH9	pH10	pH11
BPT-05	2.5	4.8	1.2	0.27	0.14	0.57
BPT-06	2	No Growth	No Growth	0.34	------	-----
BPT-08	4	No Growth	No Growth	0.23	------	-----
BPT-20	1.25	1	1.10	0.55	0.69	0.63
BPT-23	2	3	17.5	0.34	0.23	0.039
BPT-25	1	No Growth	No Growth	0.69	------	-----

* *Bacillus amyloliquefaciens* (BPT- 05), *Bacillus cereus* (BPT-06), *Staphylococcus arlettae* (BPT-08), *Bacillus licheniformis* (BPT-20), *Bacillus licheniformis* (BPT-23) and *Bacillus pumilus* (BPT-25).

a) Isolate BPT- 05

After 8 hours of lag phase, growth started in BPT-05 and it attained maximum growth after 24 hours of incubation at 37^{o}C. the growth pattern in this isolate showed that it has extreme alkaliphilic nature (Fig.3.1.14).

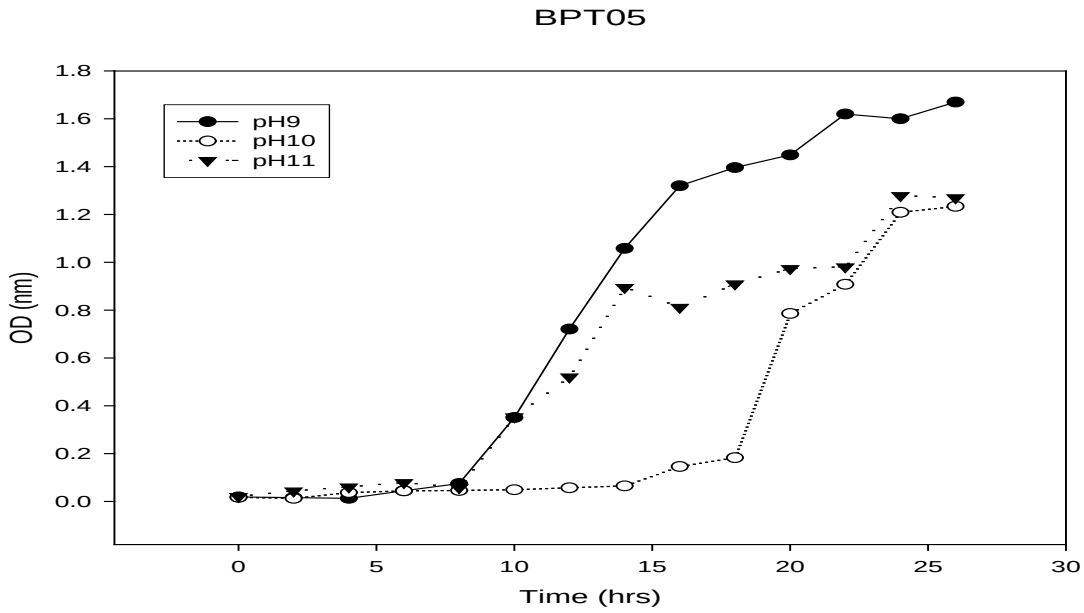


Fig.3.1.14. Growth curve of BPT-05 in Alkaliphilic media at pH 9,10, and 11.

b) Isolate BPT-06

This isolate did not show significant growth at pH 10 and pH 11 but the growth was observed after 25 hours of incubation on pH 9 (Fig.3.1.15).

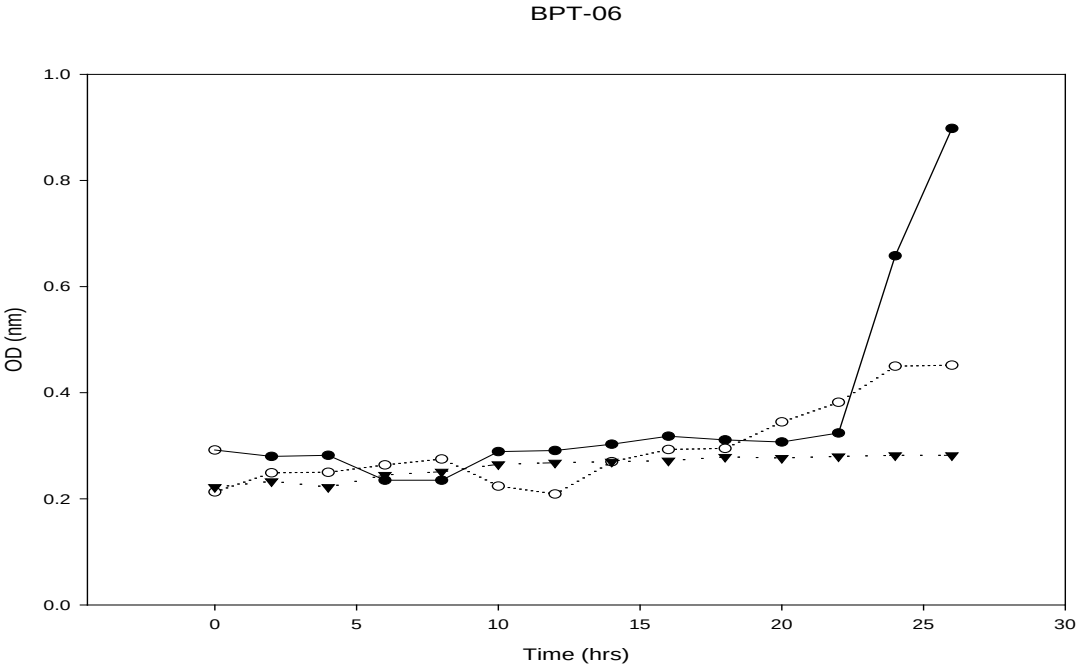


Fig.3.1.15. Growth curve of BPT-06 in Alkaliphilic media at pH 9,10, and 11.

c) Isolate BPT-08

Maximum growth was observed in BPT-08 at pH 9 after 23 hours of incubation , but no considerable growth was observed on pH 10 and pH 11 (Fig.3.1.16).

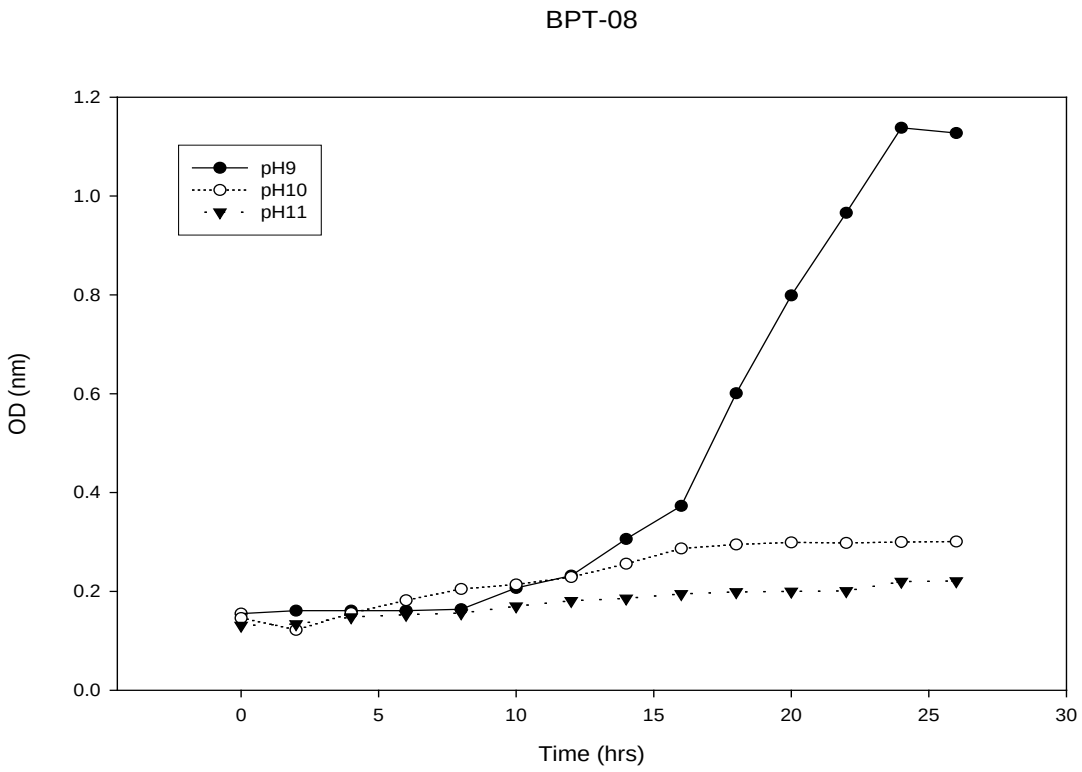


Fig.3.1.16. Growth curve of BPT-08 in Alkaliphilic media at pH 9,10, and 11.

d) Isolate BPT-20

Growth started after 24 hours of incubation at pH 9, rapid growth was observed at pH 10 and pH 11, where maximum growth was observed after 22 and 23 hours of incubation respectively (Fig.3.1.17).

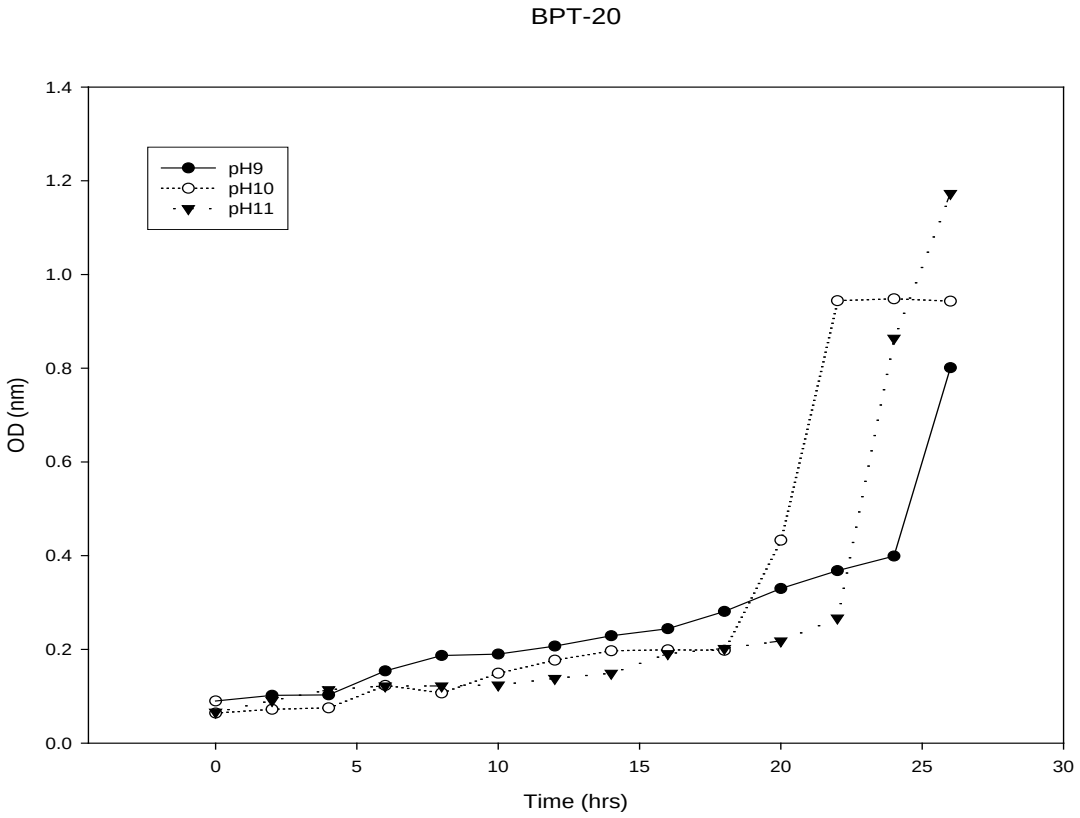


Fig.3.1.17. Growth curve of BPT-20 in Alkaliphilic media at pH 9,10, and 11.

e) Isolate BPT-23

Isolate BPT-23 had lag phase of 6 hours when grown at pH 9.0, and it attained maximum growth after 6-8 hours of incubation at pH 9, 10 and 11 respectively (Fig.3.1.18).

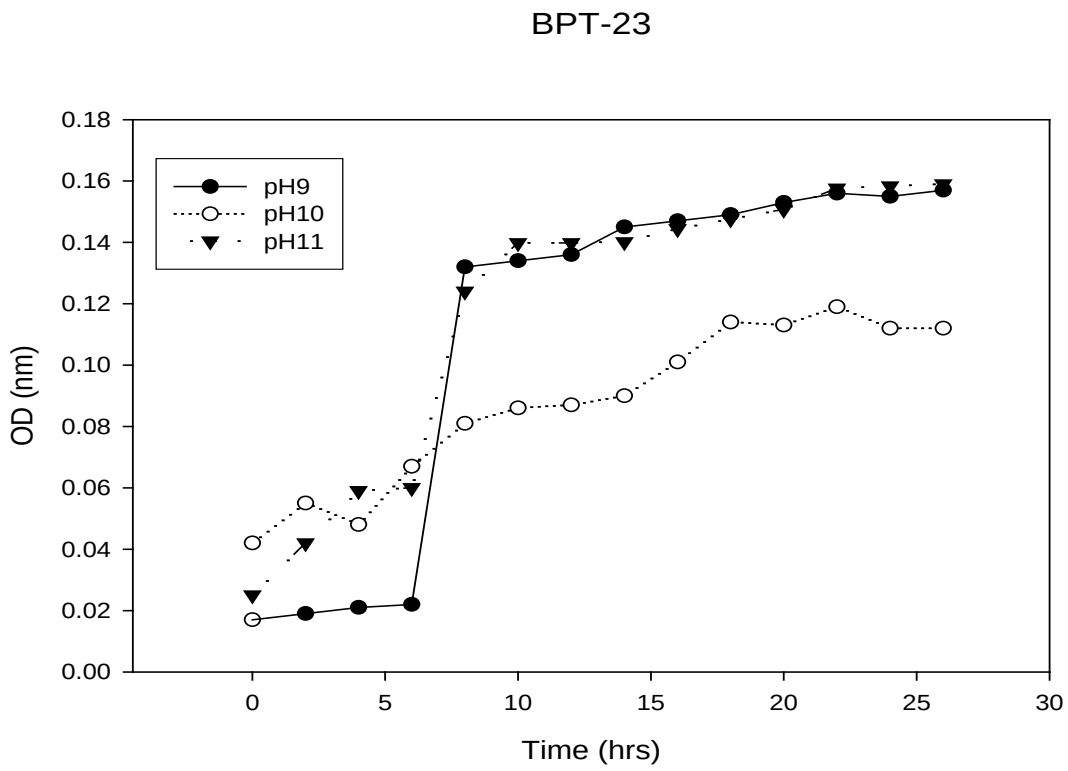


Fig.3.1.18. Growth curve of BPT-23 in Alkaliphilic media at pH 9,10, and 11.

f) **Isolate BPT-25**

Isolate BPT-25 grew at its maximum after 23 hours of incubation, having a prolonged lag phase of 22 hours. It did not showed any growth on pH10 and 11 (Fig.3.1.19).

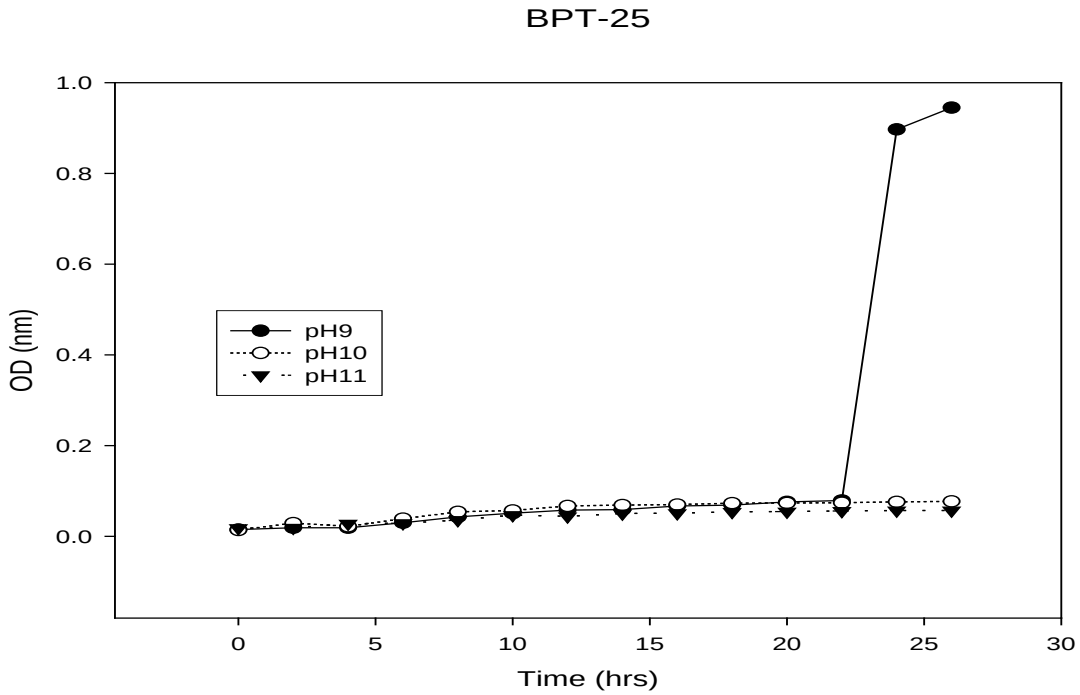


Fig.3.1.19. Growth curve of BPT-25 in Alkaliphilic media at pH 9,10, and 11.

3.1.3. Collection of metagenomic soil samples

Three soil samples were taken in sterile falcon tubes from NIBGE Canal bank (S_A), NIBGE cotton field (S_B) and NIBGE Coal Heap (S_C) to carry out metagenomic studies of protease gene (Fig.3.1.20).

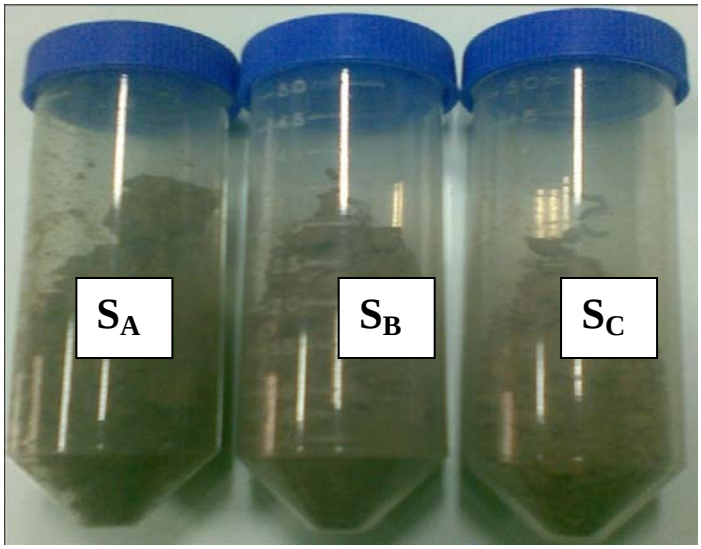


Fig. 3.1.20. Soil samples (S_A, S_B and $S_{C)}$ collected in falcon tubes from different locations at NIBGE.

3.1.3.1. Physico-chemical analyses of soil samples

Physicochemical analyses of soil samples were carried out for Na^+ and K^+,which showed that soil sample S_C had the highest content of these ions as compared to the other two samples. The soil sample $S_{C,}$ taken from the vicinity of the 20 tonnes coal heap in NIBGE, was acidic in nature while S_A was neutral and S_B was basic in nature. The acidic nature of the S_C sample was due to the accumulation of the acidic components (mostly sulfuric acid) in the surrounding of the heap during coal biodesulfurization process. Electrical conductivity (EC) of the samples ranged from 1.11-2.60. All the observations pertaining to different samples are given in Table 3.1.6.

Table. 3.1.6. Physico chemical analysis of environmental soil samples

Sample code	**Temperature (ºC)**	**Conductivity (mS/ml)**	**pH**	**Na^+ (mg/kg)**	**K^+ (mg/kg)**
S_A	25	1.11	8.56	535	568
S_B	25	0.37	7.34	470	450
S_C	25	2.60	3.00	3780	2716

3.1.3.2. Analysis of other metals present in soil samples

Soil samples were tested for different metals using atomic absorbtion spectrophometer, the results are tabulated in Table 3.1.7.

Table 3.1.7. Analysis of other metals for metagenomic soil samples.

Sample Code	Cu %	Fe %	Zn %	Ni %	Ag %	Mn%	Mg%
S_A	0.1215	0.578	0.340	0.5955	0.7645	0.4005	0.2256
S_B	0.0945	0.4883	0.5768	0.57	0.00	0.3896	0.1984
S_C	0.136	0.6190	0.4637	0.27	0.00	0.2575	0.1112

3.1.3.3. Protease activity assay

Formation of clear hollow zone around each soil sample showed that the microorganisms present in these soil samples contained active proteases, having the ability to digest casein agar media (Fig.3.1.21).

Positive protease activity led us to conduct further studies with these soil samples for the amplification of protease gene.

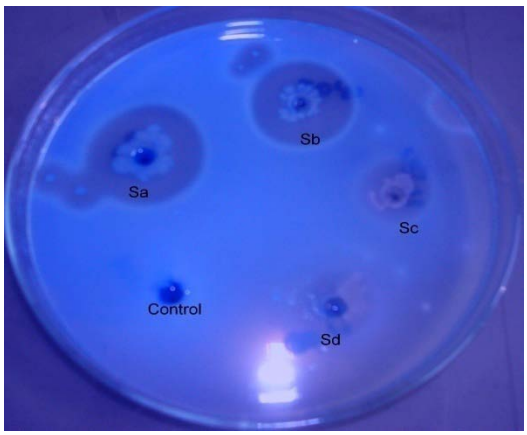


Fig. 3.1.21. Protease activity assay on metagenomic soil samples along with control.

3.2. MOLECULAR STUDIES FOR RECOVERY OF PROTEASE GENE USING SPECIFIC PRIMERS

3.2.1. Isolation of DNA from bacterial isolates

Genomic DNA from all bacterial isolates were isolated and run on 1% agarose gel. These DNAs were intact as they showed bands in a compact form and no shearing of DNA was observed (Fig.3.2.1). There were no bands for RNA or other contaminants on gel when viewed under ultra violet light, which was indicative of DNA purity.

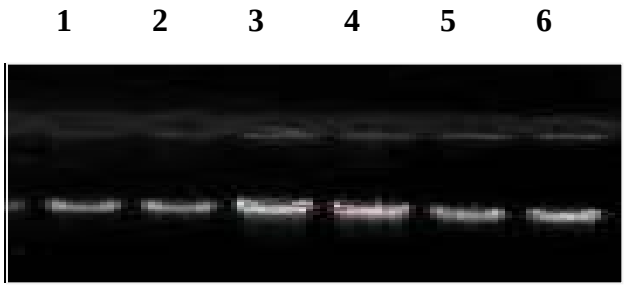


Fig.3.2.1. DNA isolation from bacterial isolates
Lanes: 1, BPT-05; 2, BPT-06; 3, BPT-08; 4, BPT-20; 5, BPT-23; 6, BPT-25.

3.2.2. DNA isolation from metagenomic soil sample

Similarly, DNA's from metagenomic soil samples were isolated using direct DNA isolation protocol. DNA isolated from soil samples are shown in Fig 3.2.2

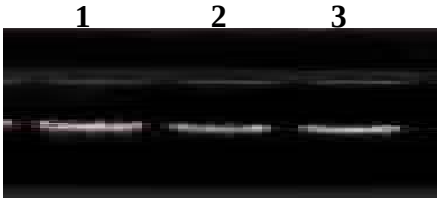
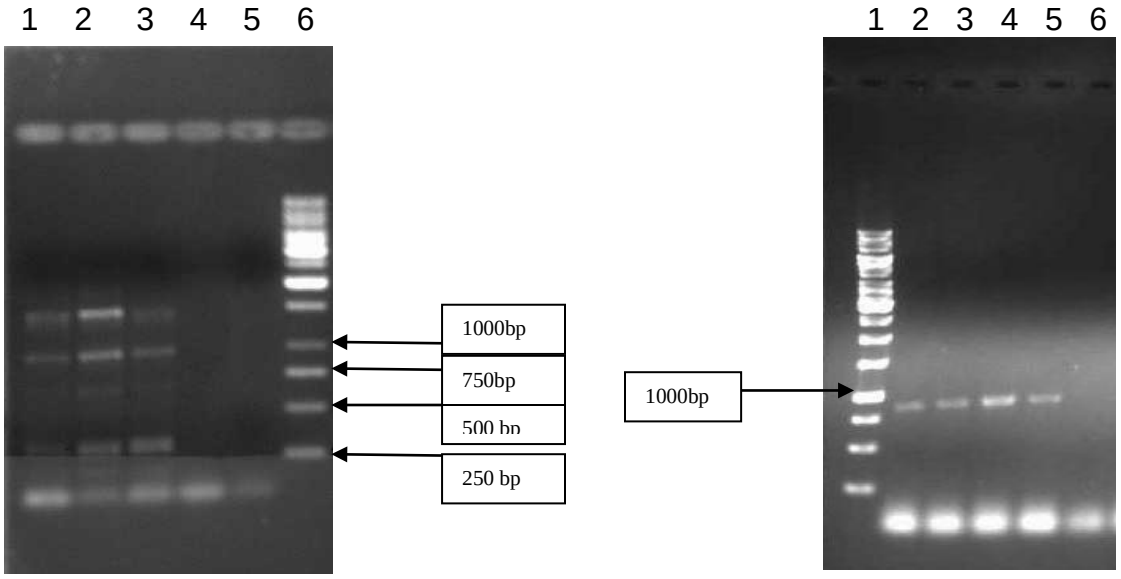


Fig. 3.2.2. DNA isolation from soil samples
Lanes: 1, S_A; 2, S_B; 3,S_C

3.2.3. Protease gene amplification of DNA isolated from bacterial isolates

Protease gene amplification was carried out for all bacterial isolates (Fig 3.2.3 , A & B). Out of these six isolates, BPT-05 and BPT-08 showed good amplification for protease gene with primer set F1&R1 and also with C1 and C2. Multiple bands for these genes were observed which were recovered by gel extraction. Similarly, protease gene for isolate BPT- 06, BPT-20, BPT-23 and BPT-25 were amplified using B1&B2 primer set. The size of amplified gene was in the range of 950-1100bp in length for these isolates. The amplified genes were designated codes as PG05, PG06, PG08, PG20, PG23 and PG25, respectively, obtained from respective isolates.



(A) **(B)**

Fig.3.2.3 (A). PCR amplification of protease genes from BPT-05 & 08
Lane: 6, 1kb ladder
Lane: 5, Negetive control
Lane: 1, 2; (PG05), 3; (PG08), Amplified region of protease gene of with F1 & R1 primers.

Fig.3.2.3 (B). PCR amplification of protease genes in BPT-06 , 20, 23,25
Lane: 1, 1kb ladder

Lane: 6, Negetive control

Lane: 2; PG06, 3; PG20, 4;PG23, 5;PG25 Amplified region of protease gene with B1 & B2 primers.

3.2.4. PCR amplification of protease gene from metagenomic DNA isolated from soil samples

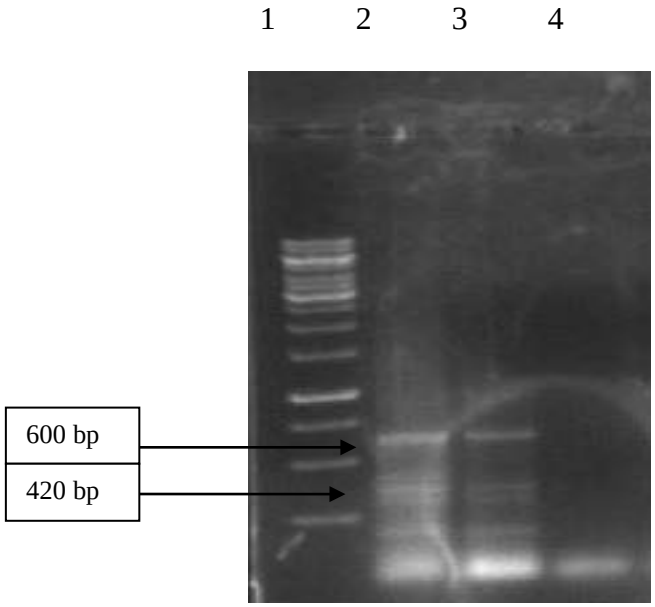


Fig. 3.2.4. PCR amplification of protease genes in S_C

Lane: 1, 1kb ladder

Lane: 4, Negetive control

Lane: 2,3; Amplified region of protease in S_C gene (PGS_C) with B1 & B2 primers.

Soil samples were processed for protease gene amplification and out of the three soil samples the amplification was observed only in S_C. (Fig.3.2.4). Multiple bands were observed in this case which were recovered through gel extraction method. The amplified gene was designated PGSc code.

3.2.5. Cloning of protease gene

Cloning of protease gene from DNA of all six isolates and metagenomic DNA were carried out using standard laboratory protocols. These clones were identified by the presence of

blue white colonies on LB plates with Ampicillin/IPTG/X-Gal. Respective white colonies from different plates were picked up and grown in LB liquid media. Plasmids were isolated and clones with protease gene as insert were selected for further processing.

3.2.6. Protease gene sequencing

Cloned protease gene and PCR fragments of the respective samples were then subjected to sequencer for sequencing. Clones were sequenced using M13 primer and PCR fragments were sequenced using Protease gene specific primers for each sample. Sequencing was also carried out commercially for the confirmation of sequenced protease gene. Commercial sequencing of PCR products from all these samples gave very good results (Appendix I). But DNA sequencing of PG-06 was unsuccessful both in our lab and also commercially. So it was not processed further.

3.3. BIOINFORMATICS STUDIES ON PROTEASE GENE

Commercial sequencing results of PCR fragments of protease genes were then further used for sequence analysis using various bioinformatics tools.

3.3.1. DNA sequence analysis

3.3.1.1. Sequence homology searches of sequenced protease genes using BLAST algorithm

Blast program of NCBI gives the homology matches with protease gene of our samples. Complete blast results are given in Appendix II. Main results are tabulated in Table 3.8. Almost all sequences show very good homology with other proteases, above 90% sequence similarity for all species except for PG-23 which shows 68% sequence identity.

Table. 3.3.1. Blast results of sequenced protease genes from different bacteria and metagenomic sample.

Gene code	Organism	Base pair size	BLAST Hit result	Homology %	Gap%	E-Value
PG-05	*B. amyloliquefaciens*	1101	*Bacillus licheniformis* strain RG2 keratinolytic protease gene	94	1	0.0
PG-08	*S. arlettae*	1156	*Bacillus licheniformis* strain YP1A protease gene	93	3	0.0
PG-20	*B. licheniformis*	819	*Bacillus pumilus* strain organic solvent tolerant protease gene	99	0	0.0
PG-23	*B. licheniformis*	950	*Bacillus pumilus* strain TMS55 alkaline serine protease precursor gene	68	1	5e-61
PG-25	*B .pumilus*	810	*Bacillus pumilus* strain 115b organic solve tolerant protease gene	98	0	0.0
PGS_C	Soil Metagenomic DNA	429	*Bacillus pumilus* strain 115b organic solvent tolerant protease gene	91	0	1e-40

3.3.1.2. Multiple sequence alignment of sequenced protease genes

Multiple sequence alignments between the sequenced protease genes as well as with protease gene sequences from other organism were constructed using CLUSTALX software. The alignment file was observed and modified through Bioedit software . Multiple sequence alignment between sequenced protease gene of five bacterial isolates and metagenomic protease gene show considerable sequence similarity. Multiple sequence alignment between our protease sequences and protease sequences from other homologous species was carried out to get the understanding of an ancestral relationship between these species. Comparison of these sequences (15 sequences) shows considerable sequence similarity among them. The result of Clustal X presenting multiple sequence alignments is presented in Appendix III.

3.3.1.3. Phylogenetic analysis

Phylogenetic tree was constructed from the results obtained from multiple sequence alignment, which shows that PG05, PG08, PG20, PG23, PG25 and PGS_C protease gene sequences have common ancestors and are homologous to each other according to evolutionary point of view (Fig.3.3.1 (A & B)).

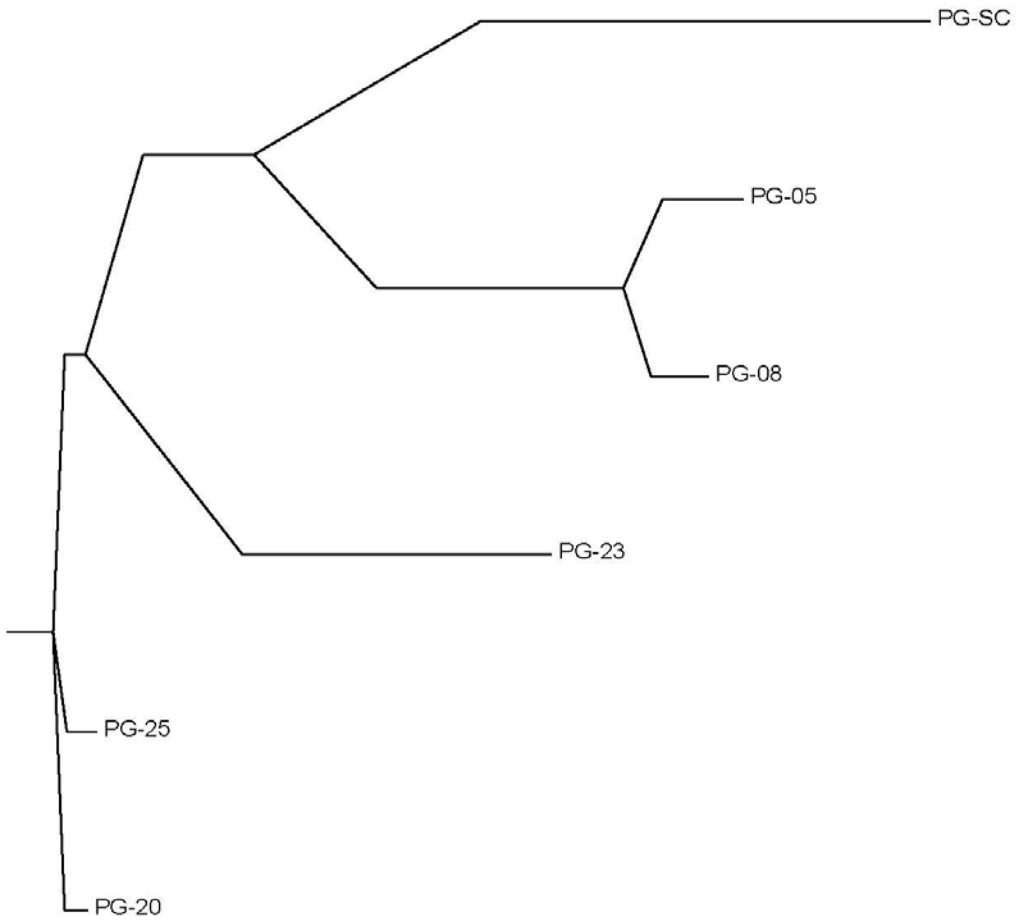


Fig3.3.1 (A). Phylogenetic tree of sequenced protease genes constructed based on multiple sequence alignment between these genes.

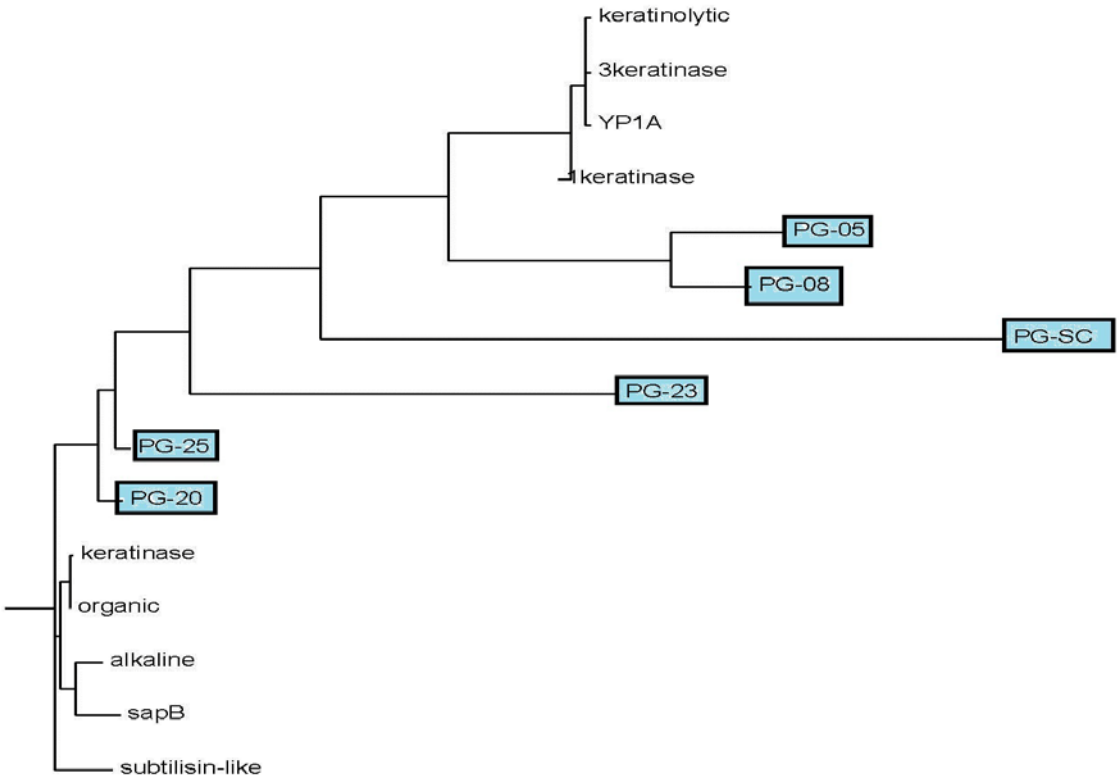
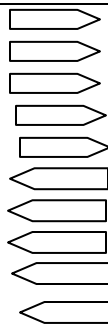
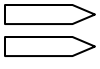
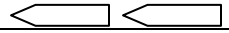
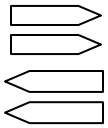


Fig.3.3.1. (B) The inferred relationship, based on protease gene sequence, isolated in this study (in blue boxes) to protease gene from other bacteria.

3.3.1.4. Open reading frames

Regions of DNA that encode proteins are first transcribed into messenger RNA and then translated into protein. By examining the DNA sequence alone we can determine the sequence of amino acids that will appear in the final protein. Once a gene has been sequenced it is important to determine the correct open reading frame (ORF). Every region of DNA has six possible reading frames, three in each direction. The reading frame that is used determines which amino acids will be encoded by a gene. Typically only one reading frame is used in translating a gene (in eukaryotes), and this is often the longest open reading frame. Once the open reading frame is known the DNA sequence can be translated into its corresponding amino acid sequence.

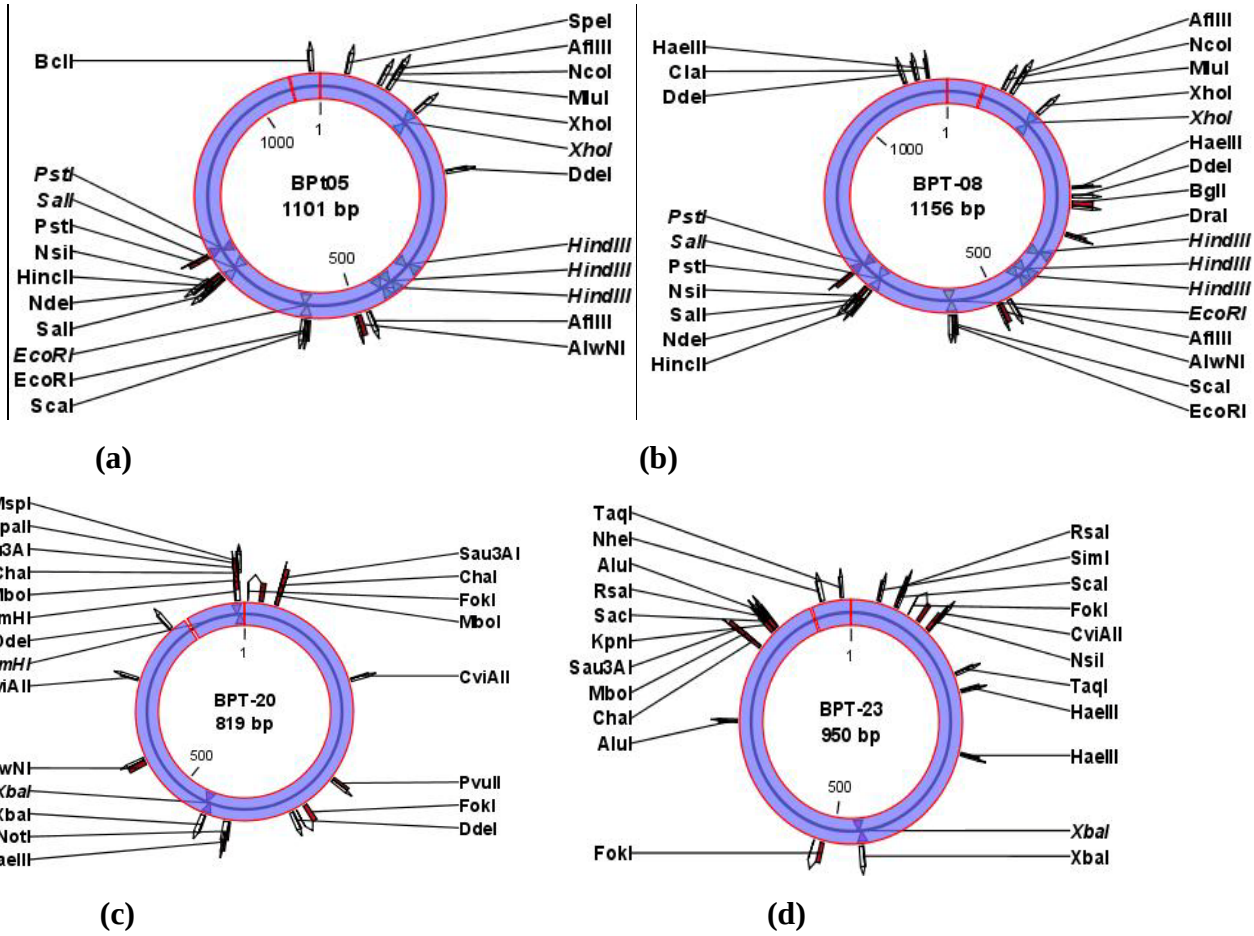
Table.10. Screening of Open Reading Frames in Protease genes from various species selected for this study.

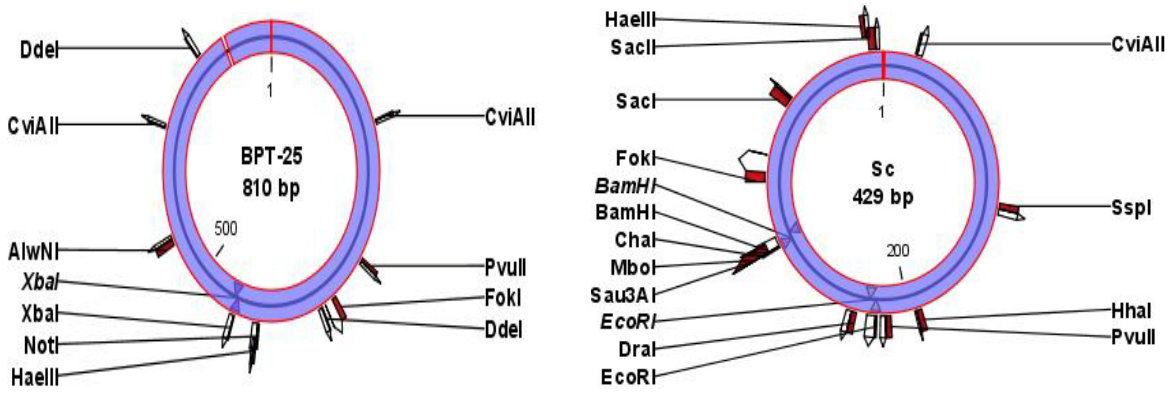
S.No	Source Code	No.of ORFs with frame No.	Orientation of ORFs	No. of amino acid in ORFs	ORF cognitor results
1.	**PG-05**	+1,+1		61+41	Novel
		+2 ,+2		165 + 58	subtilisin like serine protease gene
		+3		60	Novel
		-1		39	Novel
		-1		62	Molybdopterin biosynthesis enzyme
		-3		38	Novel
2.	**PG-08**	+1		46	Novel
		+2		62	Subtilisin-like serine proteases
		+2		56	YJR138w
		+3		50	Biotin carboxylase
		+3		36	Novel
		-1		91	Methyl-accepting chemotaxis protein
		-1		34	PA1701
		-2		62	Molybdopterin biosynthesis enzyme
		-3		65	Predicted acyltransferases
		-3		96	Novel
3.	**PG-20**	+1		163	Subtilisin-like serine proteases
		+2		43	Predicted drug exporters of the RND superfamily
		-2		55	Novel
4.	**PG-23**	+1		37	Subtilisin-like serine proteases
		+1		44	Novel
		+3,+3		58+35	Novel
		-3,-3		80+34	Novel
5.	**PG-25**	+1		43	Predicted drug exporters of the RND superfamily
		+3		162	Subtilisin-like serine proteases
		-3		55	Novel
6.	**PGS_C**	+1		48	YLR271w
		+3		45	novel
		-2		39	novel
		-3		58	Type I restriction-modification system methyltransferase subunit

Open reading frames were obtained using ORF finder programe of NCBI, the results are tabulated in Table.3.3.2. Which shows that most of the open reading frames in all protease genes were novel, a few show homology with subtilisin like serine protease , which also confirms the nature of these sequenced genes. A few open reading frames show homology with different other proteins.

3.3.1.5. Restriction map

Restricition maps were drawn using Gentle software, indicating various restriction sites of all sequenced protease genes for various restriction enzymes. Restriction map results are presented in fig. 3.3.2.





(e) **(f)**

Fig.3.3.2. Restriction maps of the protease gene for (a) PG-05, (b) PG-08, (c) PG-20, (d) PG-23, (e) PG-25, (f) PG-S_C shows various restriction sites for different restriction enzymes at different positions.

3.3.1.6. GC content

GC composition of the respective protease gene was calculated. The results are tabulated in Table.3.3.3.

Table 3.3.3. Percentage composition of the GC content presents in sequenced protease genes.

S. No	GENE ID	Length (bp)	No. of G or C bases	GC composition (%)
1.	**PG-05**	1101	559	50.8
2.	**PG-08**	1156	582	50.3
3.	**PG-20**	819	375	45
4.	**PG-23**	938	446	47.5
5.	**PG-25**	810	365	45.1
6.	**PG-S_C**	429	182	42.4

3.3.1.7.Nucleotide sequence statistics

Nucleotide sequence statistics are generated using CLC Bio main workbench. The statistics are presented in table 3.3.4.

Table 3.3.4. Sequence information statistics of protease genes from bacterial isolates and metagenomic DNA.

Information	**PG05**	**PG08**	**PG20**	**PG23**	**PG25**	**PGS_c**
Sequence type	DNA	DNA	DNA	DNA	DNA	Metagenomic DNA
Length	1101bp	1156bp	819bp	938bp	810bp	429bp
Organism	*B.amyloliquefaciens*	*S. arlettae*	*B. licheniformis*	*B.licheniformis*	*B.pumilus*	Metagenomic DNA
Description	Protease Gene	Protease Gene	Protease Gene	Protease Gene	Protease Gene	Protease Gene
Weight	356.374 kDa	374.754 kDa	264.158 kDa	300.882 kDa	261.332 kDa	138.282 kDa

Histogram of nucleotide frequencies describes how the number of a particular nucleotide that exist in a sequence. Which is show below in the Fig.3.3.3

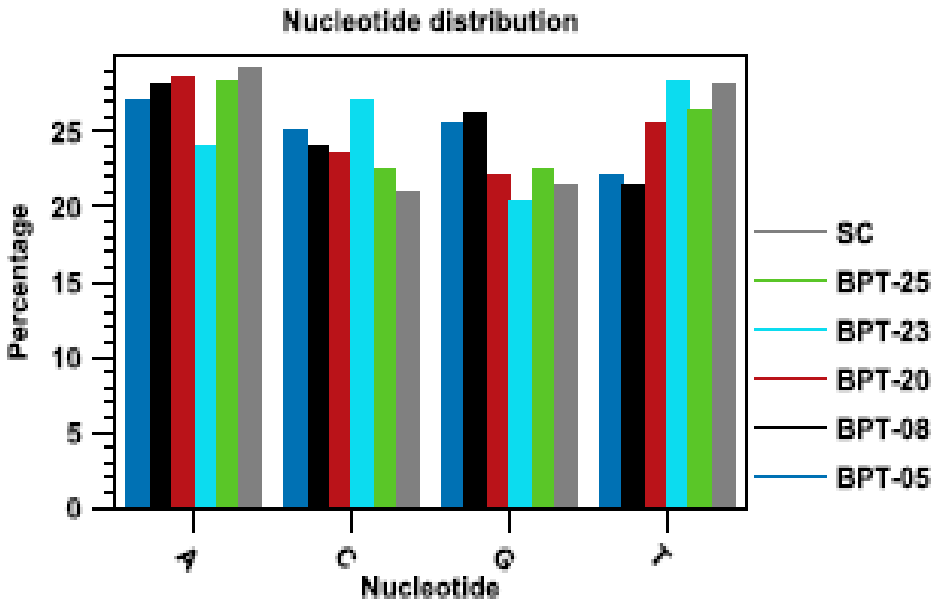


Fig. 3.3.3. Histogram of nucleotide frequencies for all sequenced genes.

3.3.1.8.Translation of DNA to protein

DNA sequences were subjected to Translate program of Expasy to get the translated Protein sequences. (http://www.us.expasy.org). which are given below.

a) PG-05

VWSYLNVASQLTSPVDAKKWASLSGVFHGFATRPLRIRRAAMSKLPLNSRFQ-CGGSGGV
PG-KVEAAAASKMRGNNV-SEAPRKNQSGSRPKLPSMLRFKGPPCWGT-FKMPRTSTKMQ
IKSRFMDTGIQASHPDLNVFGGPSFVAGEAYNTDGNGHGTHVAGTVAALDNTTGVLGVAP
SVSLYAVKVLNSSGSGSYSGIVSGIQWATTNGMDVINMSLGGASRSTAMKQAVDNAYARG
VVVVAAAGNSGSSGKHEYNWLSCKIRFCHRNWCGIL-QQQSFIFQCGSRA-SHGSWRRCI
QLLPHPTLMQHCTERQWHSPHVAGSTAMILSKHPNLSASRVRQASLTRLLICKLLLHNRK
GTDQCQ

b) PG-08

APDATSHVAAHLTILSMQRNGRHCPEFSMDLQRVH-G-GEQQCRSCRSTRDSSDSWVGRR
DARQGQSCGLTQNP-KKPSGKRLPETTKKDRGQIAPEIQPNGLPRIKAEKKQALGFKGTN
VKVAVLDTGIQASHPDLNVVGGASFVAGEAYNTDGNGHGTHVAGTVAALDNTTGVLGVAP
SVSLYAVKVLNSSGSGSYSGIVSGIEWATTNGMDVINMSLGGASGSTAMKQAVDNAYARG
VVVVAAAGNSGSSGKHEYNWLSCEIRFCHRSWVR-TLTATELHFPVVGSRA-SHGSWAQG
VYSTYQPNTYATLERNRQWLPLM-RKQQLLIFVKNIRIFQLHQVPANGLLSATSGPYLGR
LRSTKEKVLNRLSKAGPTQTSEKFQ

c) PG-20

GRTSTSIRSQQCDQGANVKIAVLDTGIHAAHPDLNVAGGASFVPSEPNATQDFQSHGTHV
AGTIAALDNTIGVLGVAPNASLYAVKVLDRNGDGQYSWIISGIEWAVANNMDVINMSLGG
PSGSTALKNAVDTANNRGVVVVAAAGNSGSSGSRSTVGYPAKYDSTIAVANVNSNNVRNS
SSSAGPELDVSAPGTSILSTVPSSGYTSYTGTSMASPHVAGAAALILSKNPNLTNSQVRQ
RLENTATPLGDSFYYGKGLINVQAASNYKRIRK

d) PG-23

LLPNVPLGS-RR-WQSTGC-FCL-CCTSWLE-CMGCEVPCVKLKCPPKFSLPWNSN
NWKHCWPCLLSGRSWGIALASLPAVQALNLSGHLH--WIISAIEGGVRNERYVVSMIFDS
VRSSARLKKAVYTH-KGGILVSAESGNYGSKNSRTKGGYPAKNQSKVAVLDVFRNLVPPE
SSTEIPDLDVSAPATSIFNNTPSFAHTSHTGTPLVSPCRLRETCLTFSNNQNLTYLQAHE
RCIKYSDTAW-LILLCENEPPRSPDR-LMIRNYMLARGTELTSHRHFLHPLDWENLTLPN
LIC-RHIHLSPLDYT-

e) PG-25

-VSFGGAKSFKGANVKVAVLDTGIHAAHPDLNVAGGASFVPSEPNATQDFQSHGTHVAGT
IAALDNTIGVLGVAPNASLYAVKVLDRNGDGQYSWIISGIEWAVANNMDVINMSLGGPSG
STALKNAVDTANNRGVVVVAAAGNSGSSGSRSTVGYPAKYDSTIAVANVNSNNVRNSSSS
AGPELDVSAPGTSILSTVPSSGYTSYTGTSMASPHVAGAAALILSKNPNLTNSQVRQRLE
NTATPLGDSFYYGKGLINVQAASTKGIPE

f)PG-Sc

QTVVIVMD-TVPYGKNQSLFRDSN-GP-N-WLERVLNNQIFFASAPFIKMSALQIAYDSS
LISCAKNSRSAVEFSSFKYSWATEPIPPVQAASN-GSKFEWAVANNKDVINMSLGGPSSS
RALKNAVDTANNRGVVVCAARG

3.3.2. Protein sequence analysis

3.3.2.1. Primary structure analysis and template search

Sequence homology searches for all translated protein sequences were carried out by using protein BLAST at NCBI. PG05 and PG08 showed 65% and 66% homology to subtilisin calsberg from *B. Subtilis* respectively. PG20 was 98% homologous to Keratinase precursor protein from *B.pumilus.* PG23 had 43% homology to serine alkaline protease preproprotein from *B.pumilus.* PG25 showed 98% homology to organic solvent tolerant protease of *B.pumilus.* Where as PGSc was 82% homologus to subtilisin protein from *B.pumilus.* The BLAST results are presented in Appendix V.

3.3.2.2. Multiple sequence alignment

Multiple sequence alignment both between translated proteases as well as protease from other bacteria were carried out using CLUSTAL X program. Results showed that there is a considerable sequence similarity between translated protease. Also these translated sequences had significant similarity between proteases from other bacterial species when they were aligned against them. There is found His active site motif (HGTHVAGTIAA) in PG05, PG08, PG20 and PG25. There is considerable active site conservation in active site region of these proteases. This motif is not present in PG23 and PGSc. Multiple sequence alignment results are given in Appendix VI.

3.3.2.4.Phylogenetic analysis

Phylogenetic tree was constructed using information file (.DND) obtained from multiple sequence alignment by the help of PHYLIP program to explore the evolutionary relationship among our proteases with proteases from other bacterial species. The results show that PG05, PG08, PG20, PG23, PG25 and PGSc have common ancestry. According to the results PG05

&PG08, PG20 & PG25, PG23 & PGSc lie close to each other having common ancestral genes with little divergence.

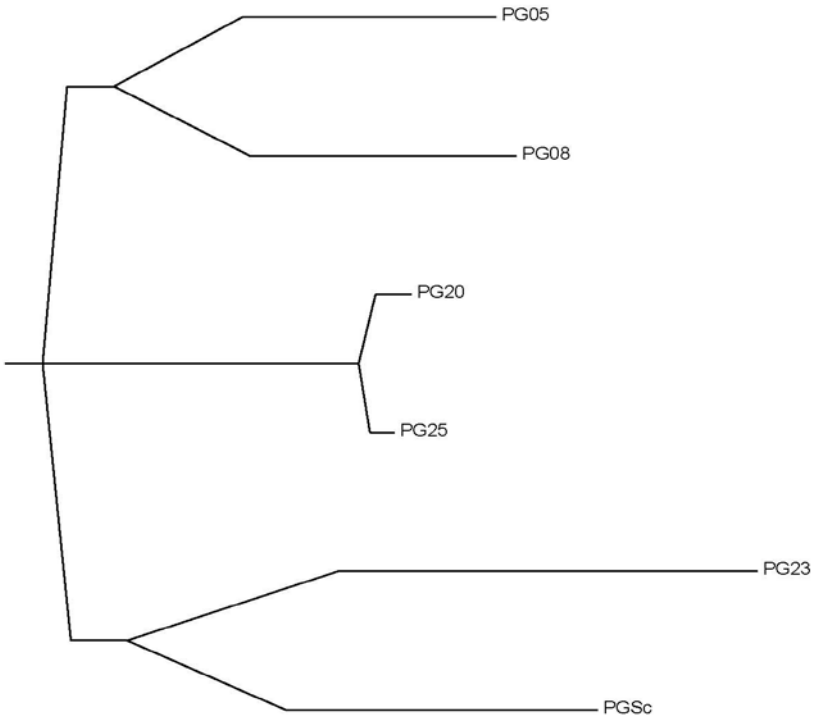


Fig.3.3.4. (A). phylogenetics tree of translated proteases.

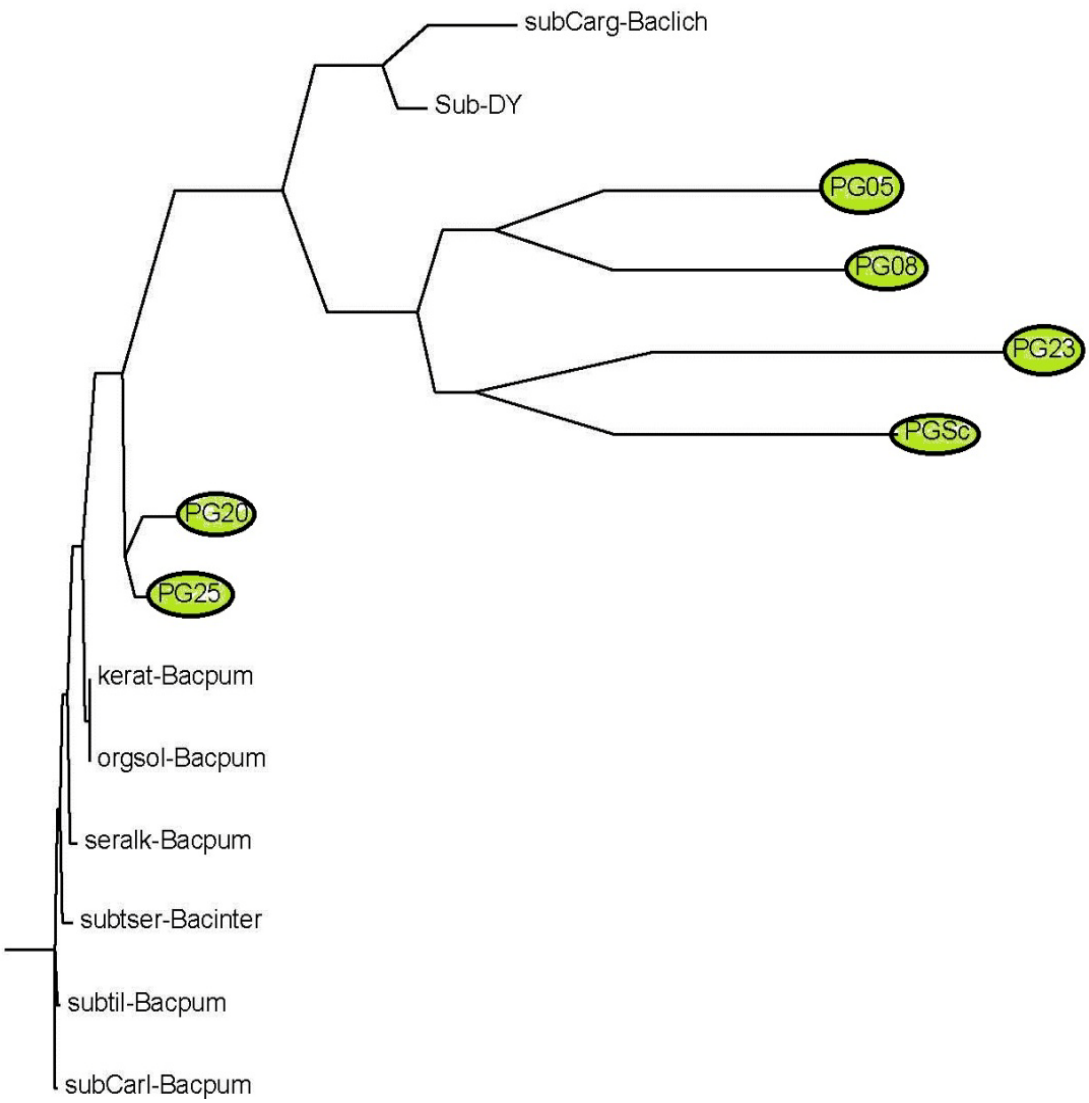


Fig.3.3.4. (B) Phylogenetic analysis of translated protease and proteases from other bacterial species. Proteases from our study are encircled in green.

3.3.2.5.Protein secondary structure prediction

Secondary structure of protein was predicted using a consensus server, which tells us about the percentage of alpha helixes, beta sheets, coils and turns present in our sequence (Table.3.3.4.). Complete secondary structure prediction results are given in Appendix VII.

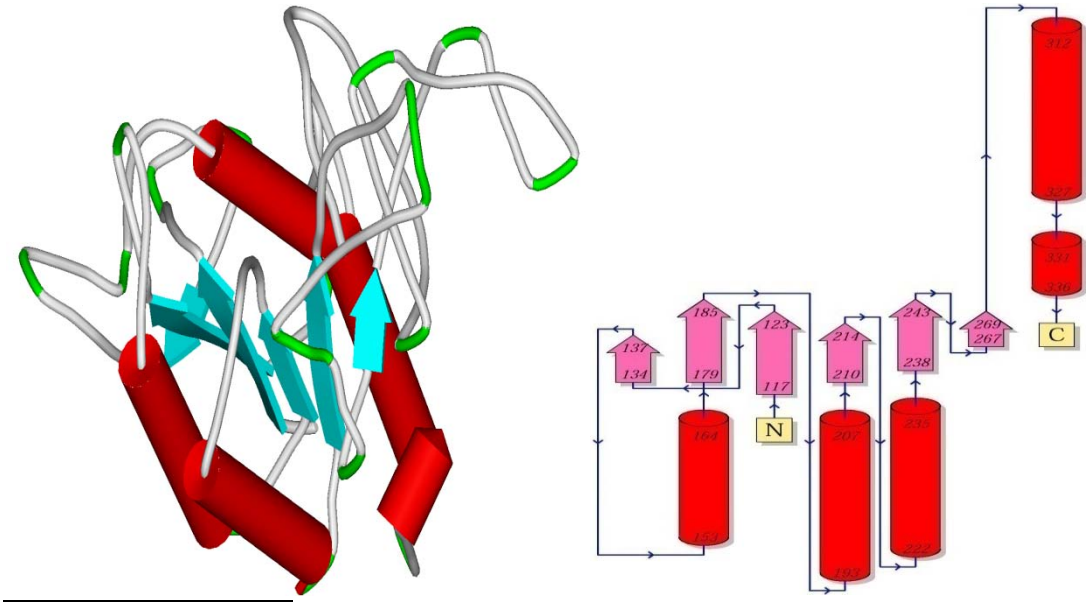
Table. 3.3.4: Protein secondary structure consensus table.

S.No	Isolate code*	Alpha helix (Hh)	%	Extended strands (Ee)	%	Random coils (Cc)	%	Ambigous states (?)	%	Total No of amino acids
1.	PG-05	66	18.33	60	16.67	208	57.78	26	7.22	360
2.	PG-08	71	18.73	65	17.15	224	59.10	19	5.01	379
3.	PG-20	59	21.61	54	19.78	158	57.88	2	0.73	273
4.	PG-23	53	17.67	44	14.67	196	65.33	7	2.33	304
5.	PG-25	26	18.98	33	24.09	76	55.47	2	1.46	268
6.	PG-Sc	50	18.66	52	19.40	161	60.07	5	1.87	137

**B. amyloliquefaciens* (PG- 05), *S. arlettae* (PG-08), *B. licheniformis (PG*-20),*B.licheniformis* (PG-23) and *B.pumilus* (PG-25),Meta genomic soil sample(PG-S_C).

3.3.2.6.Protein homology modeling and structural topology of the predicted models

PG05

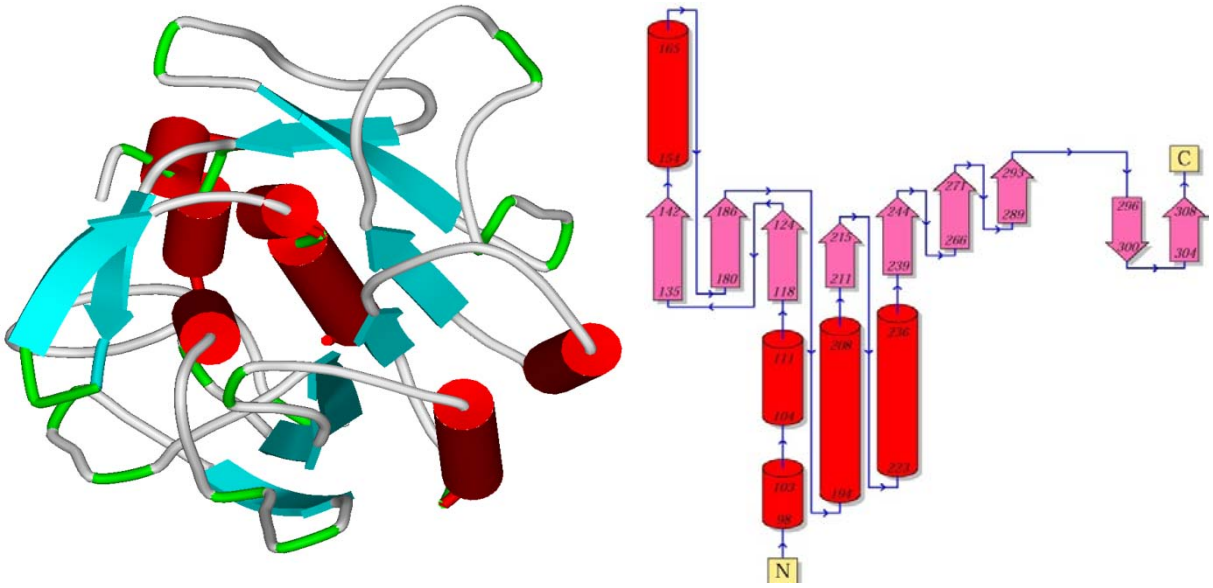


(A) (B)

Fig.3.3.5.Schematic representation of homology model along with protein topology diagram of PG05 from *B.amylilquifaciens* showing arrangement of secondary structural elements. (A) Homology model of PG05, α–helices are shown in red barrels, while β–sheets are presented with blue arrows.(B) Protein topology diagram showing the N and C terminal , 6 parallel β–sheets are sandwiched between 2 upper and three lower α–helices.

The overall tertiary structure of PG05 was visualized through 3D graphical softwares Fig. 3.3.5. Shows the schematic representation of the homology model of PG05 highlighting the arrangement of secondary structural elements. The tertiary structure comprises of α / β/ α structural topology. According to the analysis carried out by PDBsum server this protein has 1 sheet, 2 β α β units, 1 β bulge, 6 strands, 5 helices, 4 helix-helix interactions, 22 β turns and 4 γ turns.

PG08



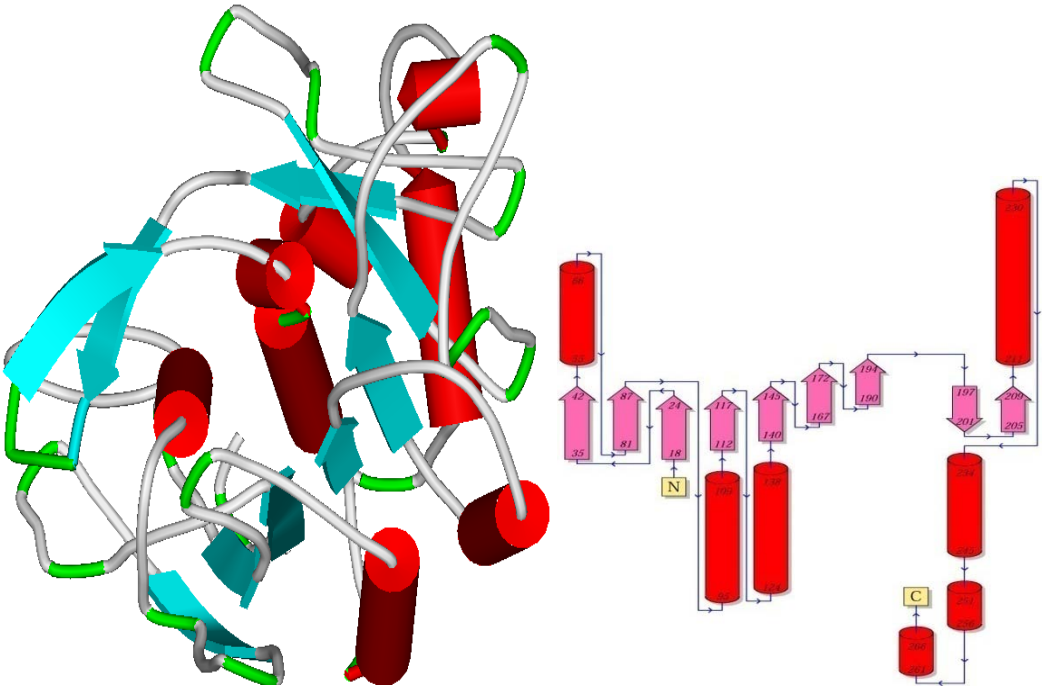
(A) (B)

Fig.3.3.6.Schematic representation of homology model along with protein topology diagram of PG08 from *S.arlettae* showing arrangement of secondary structural elements. (A) Homology model of PG08, α–helices is shown in red barrels, while β–sheets are presented with blue arrows. (B) Protein topology diagram showing the N and C terminal , 7 parallel and 2 anti parallel β–sheets are sandwiched between 1 upper and 4 lower α–helices.

The overall tertiary structure of PG08 was visualized through 3D graphical softwares Fig. 3.3.6. Shows the schematic representation of the homology model of PG08 highlighting the arrangement of secondary structural elements. According to the analysis carried out by PDBsum

server this protein has 2 sheets, C, 1 β bulge, 1β hairpin, 9 strands, 6 helices, 4 helix-helix interactions, 23 β turns and 2 γ turn

PG20

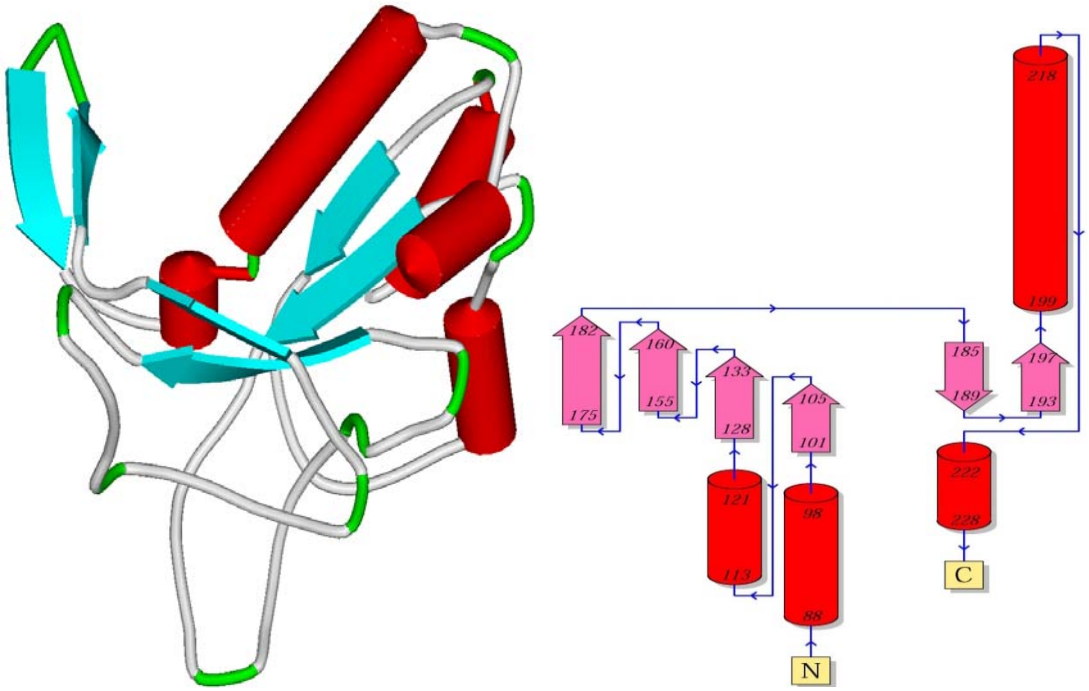


(A) (B)

Fig.3.3.7.Schematic representation of homology model along with protein topology diagram of PG20 from *B.licheniformis* showing arrangement of secondary structural elements. (A) Homology model of PG20, α–helices are shown in red barrels, while β–sheets are presented with blue arrows.(B) Protein topology diagram showing the N and C terminal , 7 parallel and 2 anti parallel β–sheets are sandwiched between 1 upper and 4 lower α–helices.

The overall tertiary structure of PG020 was visualized through 3D graphical softwares Fig. 3.3.7. Shows the schematic representation of the homology model of PG20 highlighting the arrangement of secondary structural elements. According to the analysis carried out by PDBsum server this protein has 2 sheet, 2 β α β units, 1 β bulge, 1β hairpin,9 strands, 6 helices, 4 helix-helix interactions, 23 β turns and 2 γ turns

PG23

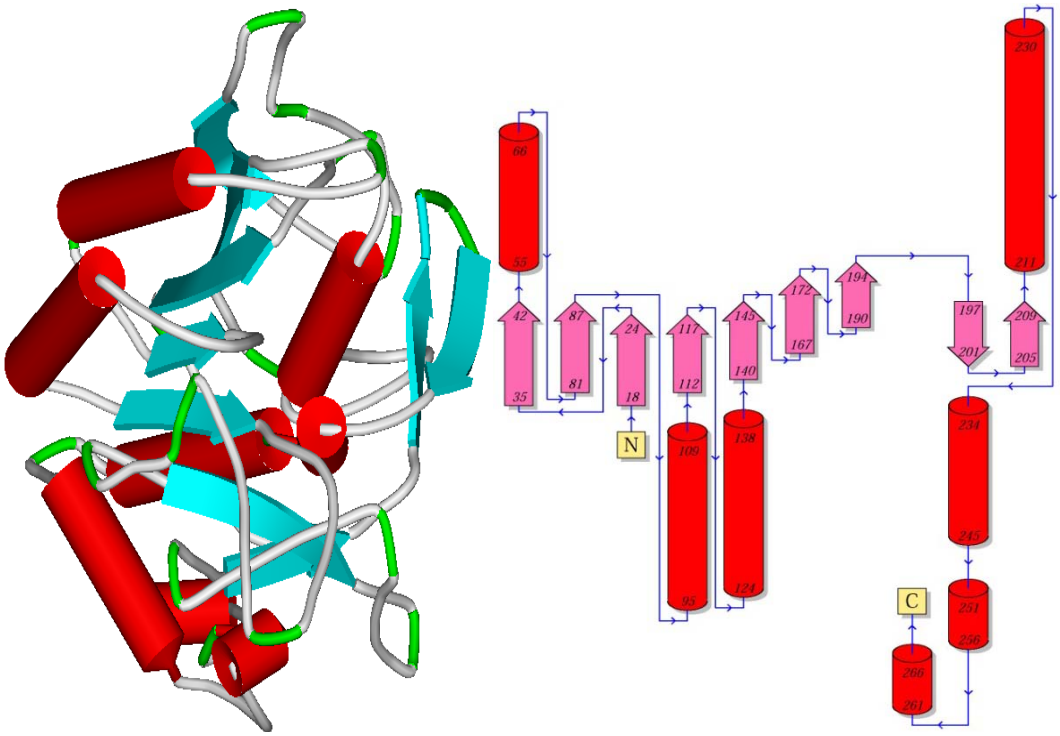


(A) (B)

Fig.3.3.8.Schematic representation of homology model along with protein topology diagram of PG23 from *B.licheniformis* showing arrangement of secondary structural elements. (A) Homology model of PG23, α–helices are shown in red barrels, while β–sheets are presented with blue arrows.(B) Protein topology diagram showing the N and C terminal , comprising of 4 parallel β–sheets linked to 2 α–helices, another 2 anti parallel β–sheets are sandwiched between 1 upper and 1 lower α–helix

The overall tertiary structure of PG23 was visualized through 3D graphical softwares Fig. 3.3.8. Shows the schematic representation of the homology model of PG23 highlighting the arrangement of secondary structural elements. According to the analysis carried out by PDBsum server this protein has 2 sheet, 1 β α β units, 1 β bulge, 1β hairpin, 6 strands, 4 helices, 2 helix-helix interactions, 16 β turns and 1 γ turn**.**

PG25



(A) **(B)**

Fig.3.3.9.Schematic representation of homology model along with protein topology diagram of PG25 from *B.pumilus* showing arrangement of secondary structural elements. (A) Homology model of PG25, α–helices are shown in red barrels, while β–sheets are presented with blue arrows.(B) Protein topology diagram showing the N and C terminal , N-terminal domain comprising of 7 parallel β–sheets linked to 2 upper and 1 lower α–helix, another 2 anti parallel β–sheets are linked to 1 upper and 3 lower α–helix in the region of C-terminal domain.

The overall tertiary structure of PG25 was visualized through 3D graphical softwares Fig. 3.3.9. Shows the schematic representation of the homology model of PG25highlighting the arrangement of secondary structural elements. According to the analysis carried out by PDBsum server this protein has 2 sheet, 2 β α β units, 1 β bulge, 1β hairpin, 9 strands, 7 helices, 6 helix-helix interactions, 23 β turns and 2 γ turns.

PGSc

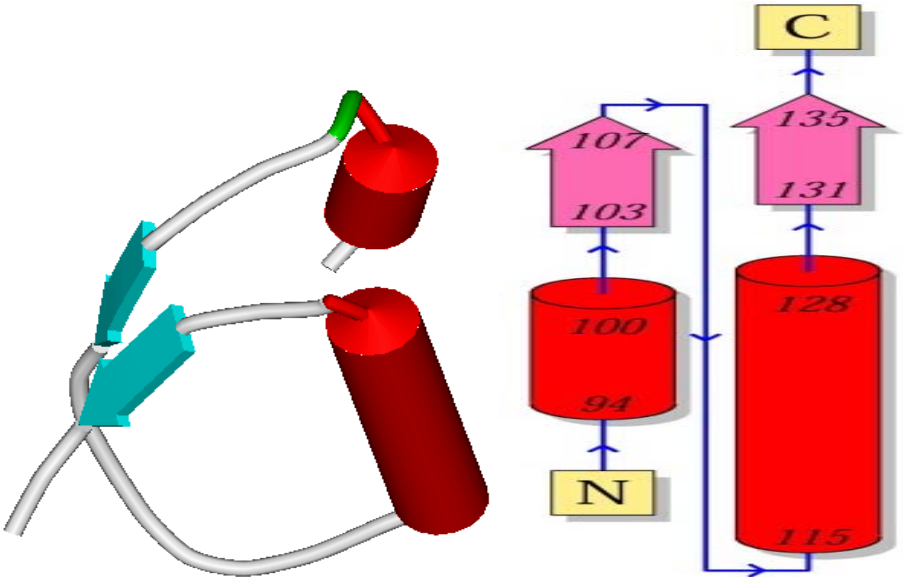


Fig.3.3.10.Schematic representation of homology model along with protein topology diagram of PGSc obtained from metagenomic sample showing arrangement of secondary structural elements. (A) Homology model of PGSc, α–helices are shown in red barrels, while β–sheets are presented with blue arrows.(B) Protein topology diagram showing the N and C terminal , comprising of 1 parallel β–sheets linked to 1 small α–helix, C-terminals also have 1 β–sheets but a large α–helix.

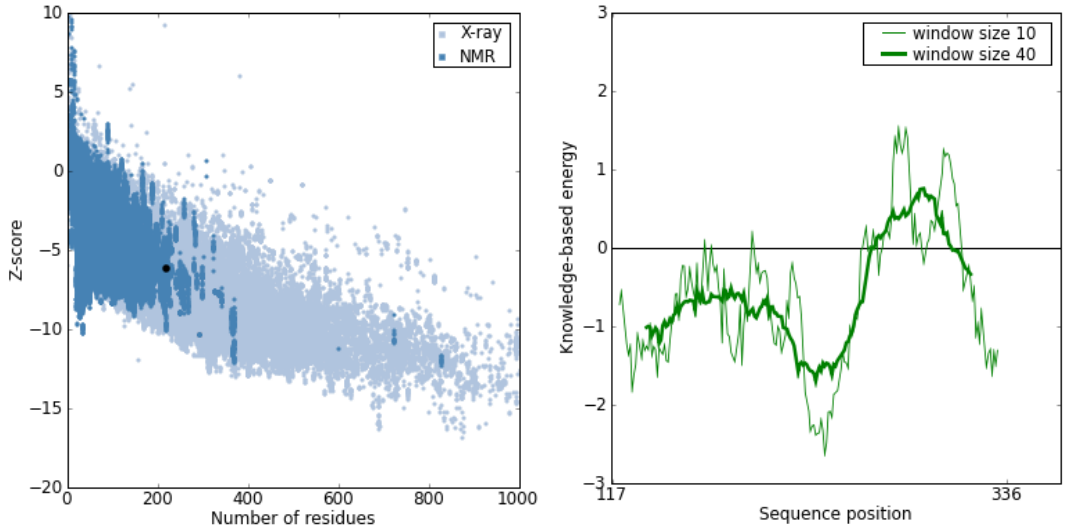
The overall tertiary structure of PGSc was visualized through 3D graphical softwares Fig. 3.3.10. Shows the schematic representation of the homology model of PGSc highlighting the arrangement of secondary structural elements. According to the analysis carried out by PDBsum server this protein has 1 sheet, 1 β α β units, 2 strands, 2 helices, 1 helix-helix interactions, 1 β turns and 1 γ turn.

3.3.2.7. Model evaluation

Ramachandran plot indicates that 88.4%, 87.8%, 85.4%, 83.6%, 84.7%, 91.9% of the main-chain dihedral angles are found in the most favored regions for PG05, PG08, PG20, PG23, PG25 and PGSc respectively. There are zero residues in disallowed region for all protein models except for

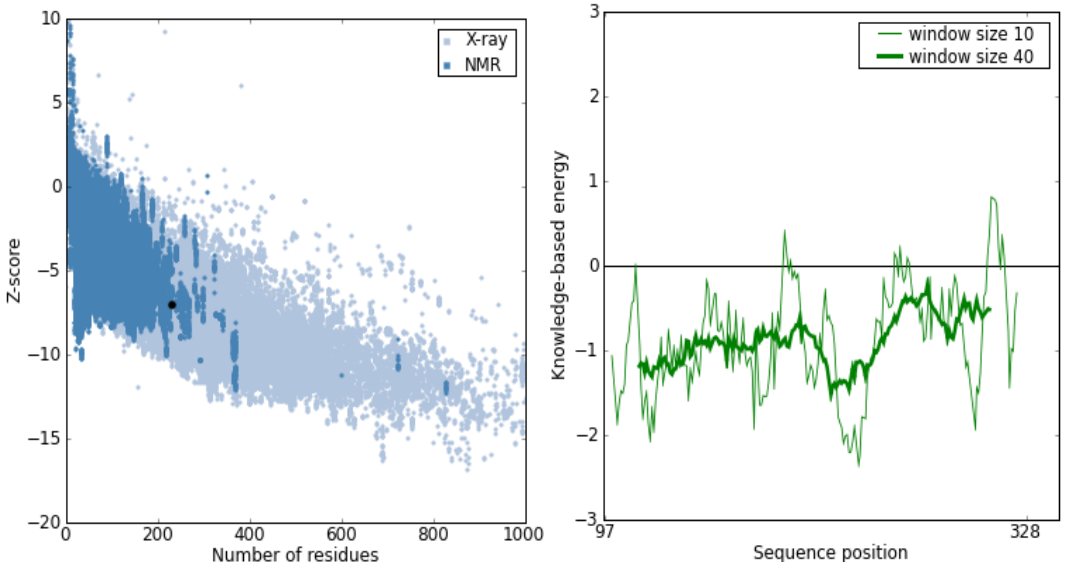
PG23 which has 0.8% residues in the disallowed region. The Ramachandran Plots of all proteins are presented in Appendix VIII.

Prosa results shows z-score values below zero for all protein models. Graphical plots below indicates the overall quality of these proteins in terms of z-values



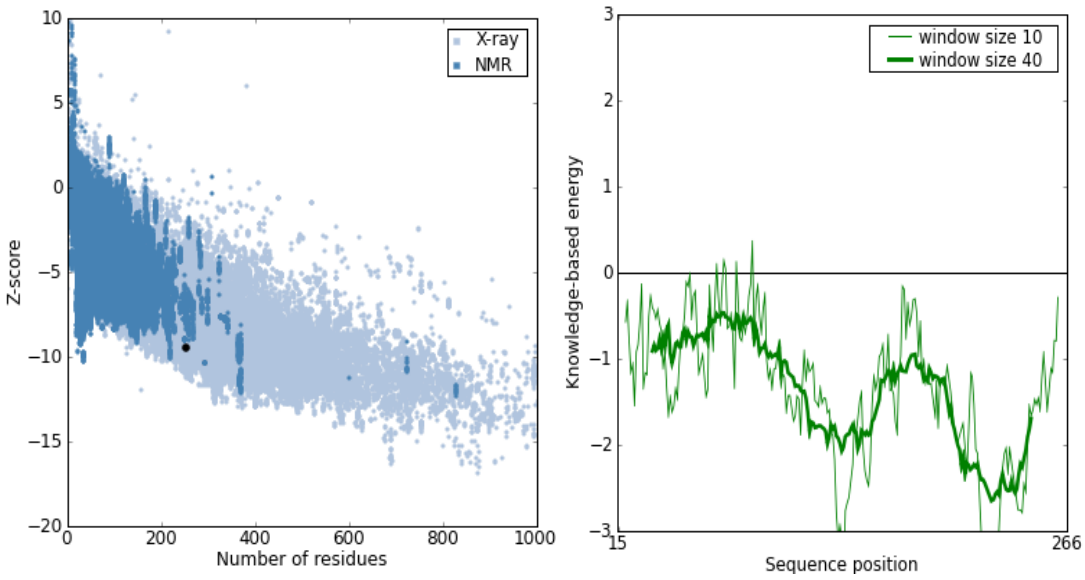
(A) (B)

Fig.3.3.11. ProSA plots. (A) ProSA z-score plot of PG05 showing the Z value < 0.
(B) PROSA-WEB energy graph of residue scores of a native protein structure for PG05



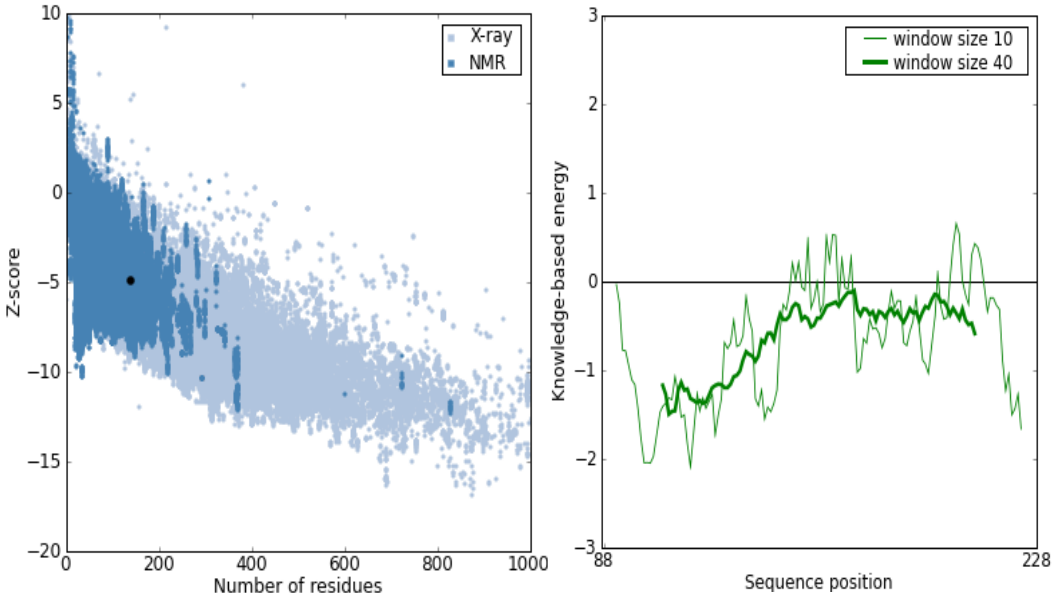
(A) (B)

Fig3.3.1.12 Plot of ProSA z-score plot of PG08 showing the Z value < 0. PROSA-WEB energy graph of residue scores of a native protein structure for PG08



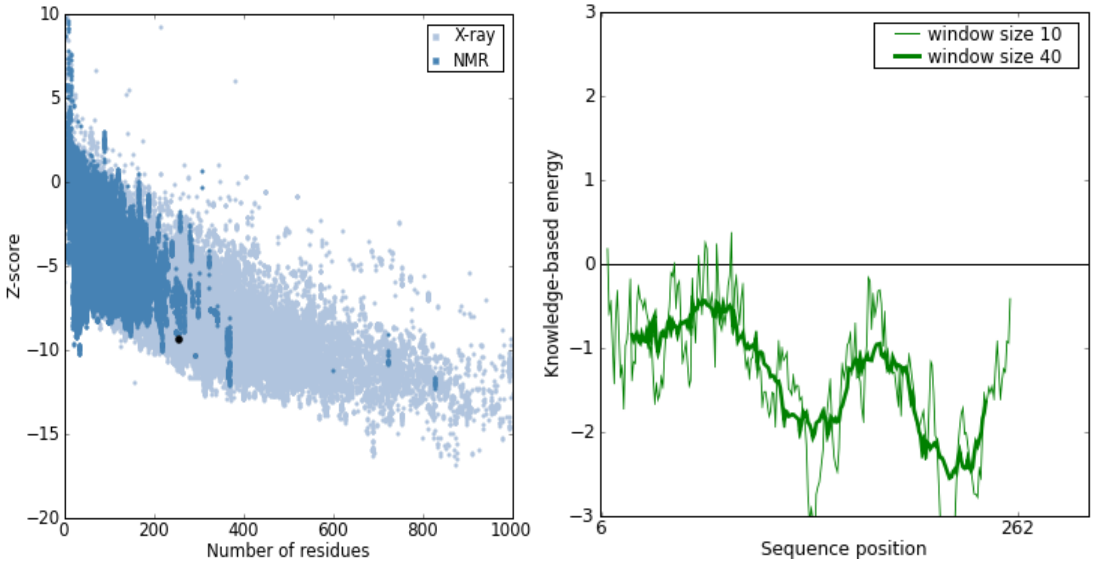
(A) (B)

Fig3.3.13.ProSa plot.(A) ProSA z-score plot of PG20 showing the Z value < 0.
(B)PROSA-WEB energy graph of residue scores of a native protein structure for PG20



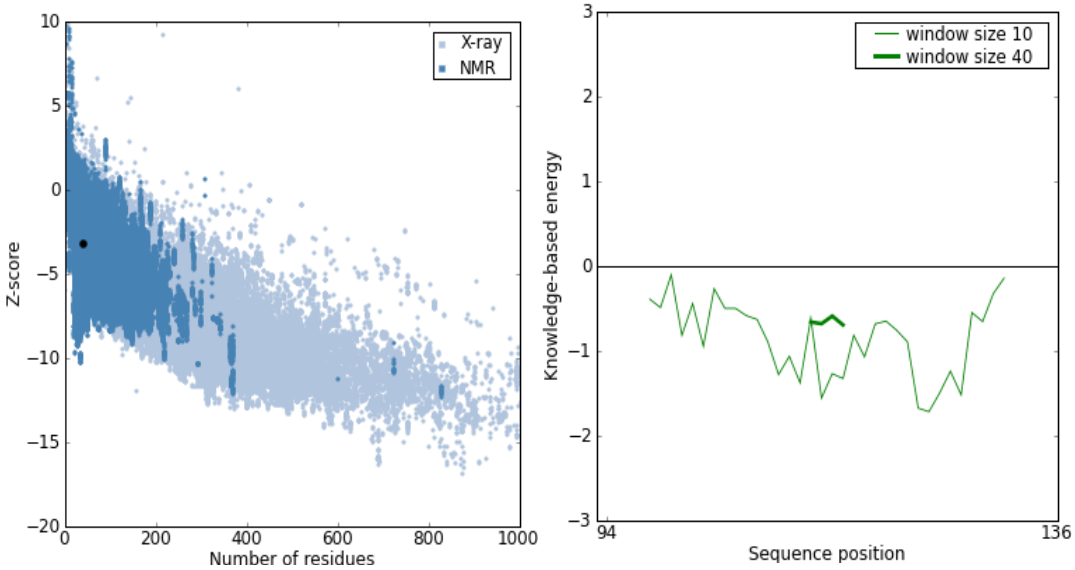
(A) (B)

Fig.3.3.14.ProSA plots. (A)ProSA z-score plot of PG23 showing the Z value < 0.
(B)PROSA-WEB energy graph of residue scores of a native protein structure for PG23



(A) (B)

Fig.3.3.15. ProSA plots. (A) ProSA z-score plot of PG25 showing the Z value < 0. (B)PROSA-WEB energy graph of residue scores of a native protein structure for PG25



(A) (B)

Fig.3.3.16. ProSA plots. (A) ProSA z-score plot of PGSc showing the Z value < 0.
(B) PROSA-WEB energy graph of residue scores of a native protein structure for PGSc

CHAPTER

4.0

DISCUSSION

4.1. MICROBIOLOGICAL STUDIES FOR PROTEASE GENE

4.1.1. Diversity , importance, global economic impact and future projections in industry

Interest in the biodiversity of extreme environment has grown over the past several years for several reasons, including the theory that such conditions were predominant on the young planet Earth. Thus early life forms may have consisted of organisms adapted to such environments. The same line of reasoning underlies the search for life forms from outside our planet. Other reasons for exploring the biodiversity of extreme environments are of more applied nature, including the use of thermostable enzymes in industrial applications (Hallberg and Johnson, 2001).

In pharma industry many proteases are used as potential drug targets for diagnostics and prognostic biomarkers. Proteases are of major focus of attention in biotechnology industry because of their effective use as biochemical reagents also for production of industrial products. This remarkable diversity in function of proteases is the direct out come of evolutionary impact of array of enzymes exhibiting a variety of sizes and shapes. The size of protease ranges from small enzymes made up of simple catalytic units to sophisticated protein processing and degradation machinaries. While discussing specificity, diversity is equally important. Protease have different way of establishing their location in cellular geography.where they operate as complex network proving additional level of interest to study of proteolytic enzymes.

Comparative genomics has paved way to understanding of the conservation , evolution and functional relevance of these enzymes. Global analysis of proteolytic system resulted in emerging patterns presenting an example of diversity and multiplicity.

From these view points it is evident that despite having proteolytic centers conservation of these proteases in all organisms , some unique proteases also play significantly specific roles in different species. (Carlos Lo´ pez-Otí and Judith S. Bond., 2008).

Proteases are known to be ubiquitous enzymes with diverse applications in many industries mainly in the detergent and food industrial sector. These industries includes beer, chill proofing, meat, cheese manufacturing, flavor development, information, baking and health product manufacturing(Yang et al., 2000). This makes proteases one of the highest value commercial enzymes (Bhosale et al., 1995).Proteases accounts for atleast a quarter of total global enzymes (layman., 1986).

Current study was carried out to study protease gene diversity in microbial life forms, also in metagenomic samples to explore variety of proteases found in indigenous soil of varied nature. Microbial proteases have all the preferred characteristics needed for biotechnological applications; this was the reason for selection of industrially important microbial protease gene for study. The current study dealt with the investigation of molecular aspects also exploration of structural and functional features of protease gene to get more better understanding of roles of proteases produced by them.

4.1.2. Pure cultures and of metagenomic samples as resource materials for the study of protease enzyme.

Culture dependent methods rely on the chemical composition of microorganisms in the environment. Bacteria change their chemical composition substantially to accommodate environmental fluctuations. Thus, when comparing bacteria on the basis of some chemical components, it is important that variation observed is a result of genetic

differences and not due to an environmental effect. Of the various components used for taxonomy only chromosomal DNA and RNA are unaffected by growth conditions. The amount of these molecules will fluctuate with growth rate, but the composition is invariant. Thus the nucleic acids offer the only standard molecules by which the widest range of microorganisms (and higher eukaryotes) can be compared and classified. Constitutively synthesized proteins are also very useful in this respect but individual proteins may not be distributed universally (Brian and Fergus 1986).

At present biotechnology is in need for finding novel genes which can be applicable in various industrial processes and development of genetically modified organisms. Identification, isolation and cloning for novel genes with rapid rate is the main vigor behind the development of extraordinary experimental approaches. Metagenomics aims for the discovery of novel genes and represents a new approach in genome level analysis. Metagenomic techniques explores the potential reservoir of new genes in soil and various environmental samples. Complex microbial communities (both cultivable and uncultivable) provides a affluent source of novel genes required in biotechnological applications isolated through metagenomic approaches. Metagenomics has remakablely contributed to the already existing repository of prokaryotic genes but it needs to grow more. (J. Vakhlu et al., 2008). In the present study we have applied separate techniques for DNA isolation from pure cultures and from metagemonic soil sample for harvesting novel protease genes.

In current work screening and isolation of microorganisms pesticide industry waste was done coupled with a molecular approach for confirmining the identity of the dominant bacterial isolates reveled that there was considerable microbial biodiversity in

the samples. Microbial diversity that was evident in preliminary culturing and characterizing studies was further analyzed through advanced molecular techniques. On the basis of the Gram's reaction, most of the isolates belonged to the Gram-negative group, which does not necessarily exclude the possibility that these isolates may belong to a diverse taxonomic linage. PCR amplification of 16S rDNA sequence-based bacterial identification was used for molecular study. The reason being it is most used, reliable and precise method currently in practice for taxonomic evaluation.

The 16S rDNA results extended the preliminary observations, and indicate considerable phylogenetic diversity within a closed ecological environment. But unfortunately those samples were found protease negative when they were tested on casine agar plates, that is why they were not processed further for protease gene hunt. The 16S rDNA sequence results obtained from this study were submitted to GenBank records.

4.1.3. Protease assay development for pure cultures and metagenomic samples

The aim of the present study was to profound our knowledge about exploration of protease gene using culture dependent and culture independent methods. After the failure of protease gene discovery from environmental samples, six bacterial isolates were selected from NIBGE culture collection.. For metagenomic studies soil samples were collected from different locations around NIBGE. Both bacterial isolates and soil samples gave positive results when tested for protease activity.

All the six pure extremophilic bacterial isolates i.,e, *Bacillus amyloliquefaciens* (BPT- 05), *Bacillus cereus* (BPT-06), *Staphylococcus arlettae* (BPT-08), *Bacillus licheniformis* (BPT-20), *Bacillus licheniformis* (BPT-23) and *Bacillus pumilus* (BPT-25)

were subjected to dual susbtrate plate assay for protease activity. All of these isolates were found protease positive, which was indicated by the formation of a clear zone around spot inoculation (Fig 3.1.6). These isolates were further used for the amplification and characterization of protease gene.

4.1.4. Growth studies in different media

Characterization and optimization for each factor of growth, nutritional requirement, and production yields are essential requirements before the selected strain is used for further investigation (Lee *et al*., 2002).

Bacterial isolates were employed to grow in halophilic media with different salt concentrations and alkaliphilic media with different pH values. The yield improvement of protease, in general, by any microbial system depends on the physiological, nutritional, and biochemical nature of the microbe employed, and these factors vary from organism to organism (Parakasham *et al*., 2005). The notable environmental and fermentation factors that influence metabolism-mediated production yields include pH, temperature, aeration, agitation, carbon and nitrogen sources, salt and metal ion requirement, incubation time and initial innoculum size, etc. Keeping this in view, preliminary investigations were made to understand various nutritional requirements and growth parameters of bacterial isolates obtained from NIBGE culture collection.

Generally bacterial growth comprises of four phases i.,e. lag phase, log phase, stationary phase, death phase. During the lag phase, active metabolic activity occurs involving synthesis of DNA and enzymes, but no growth. Geometric population growth occurs during the log, or exponential phase, when metabolic activity is most intense and cell reproduction exceeds cell death. Following the log phase, the growth rate slows and

the production of new cells equals the rate of cell death. This period, known as the stationary phase, involves the establishment of an equilibrium in population numbers and a slowing of the metabolic activities of individual cells. The stationary phase reflects a change in growing condition. For example, a lack of nutrients or the accumulation of waste products.

4.1.5. Growth in halophilic conditions

High salinity represents an extreme environment that relatively few organisms have been able to adapt to and occupy. Most halophilic and all halotolerant organisms expend energy to exclude salt from their cytoplasm to avoid protein aggregation 'salting out'. In order to survive the high salinities, halophiles employ two differing strategies to prevent desiccation through osmotic movement of water out of their cytoplasm.(santos et al., 2002).

Halophiles can be broken down into three different groups; instead of optimum growth temperature, the groups are based on optimum salinity. There are slight halophiles that grow at an optimum salinity 2% to 5%, moderate halophiles that grow at an optimum salinity of 5% to 20%, and finally extreme halophiles that grow at an optimum salinity of 20% to 30%. Also, some organisms are referred to as "halotolerant," meaning that the organism has the ability to grow in both hypersaline environments and non-saline environments, but saline is not required for optimum growth (dasserma 1999). The isolates understudy displayed different bahaviour at varied salt concentrations. The prolonged lag phase indicates that these organism took time to adapt themselves to extreme saline environment. Isolate BPT-05,06,08,and 20 showed maximum growth after approximately 19 hours of incubation in 3% and 6% NaCl concentrations. These isolates

did not showed growth in 9% salt concentration, BPT-25 did showed some growth in 9% salt concentration but it is not significant. Whereas BPT-23 expressed maximum growth after 19 hours and 21 hours of incubation in 3%, 6% and 9% salt concentration. From these results it can be inferred that except from BPT-23 all other isolates reported to show moderate halophilic behavior we can call these isolates as halotolerant as these isolates have adapted themselves to extreme halophilic conditions . In comparison to this BPT-23 showed extreme halophilic nature.

4.1.6. Growth in alkaliphic conditions

A wide variety of *Bacillus* species (Markland and Smith, 1971) secrete serine endoproteases into the external medium. *Bacillus* serine proteases have their best-known application in detergent powders. To best meet the alkaline conditions in detergents, serine proteases with a highly alkaline pH optimum are preferred above the subtilisins that have an optimal pH of 8.5 to 10. After a screening programe, Zuidweg *et al*. (1972) isolated an alkalophilic *Bacillus* strain (PB92) that produced a high-alkaline serine protease (PB92 protease) with unique pH optimum of 10.5 to 12. The PB92 protease performed extremely well in detergents. Amino acid sequences for the serine proteases from *Bacillus amyloliquefaciens*, *Bacillus licheniformis*, and *Bacillus subtilis* have been determined (Nedkov *et al*., 1985), and the genes from *B. amyloliquefaciens, B. licheniformis, B. subtilis, B. subtilis subsp. amylosacchariticus,* and the alkalophilic *Bacillus sp*. strain YaB have been cloned and sequenced (Kaneko *et al*., 1989).

Denizci *et al*. (2004) reported that alkaliphilic *Bacillus* strains are quite important and interesting for both academic and industrial reasons. One unit of alkaline protease activity was defined as the amount of enzyme able to produce 1µg tyrosine in 1 minute under

assay conditions *Bacillus sp*. grows very fast and formation of protease starts after 5 h of growth and reaches its maximum in 9 hours (1.93 U / mg proteins) and then begins to fall (Ward, 1985). Relation between protease production and growth is well studied and documented in literature (Parakasham, 2005).

Present results of time course study revealed that BPT-05 grew very fast when inoculated at 5% of innoculum size. Maximum growth obtained after 24 hours of incubation at 37°C in the liquid medium at pH 9.0, 10.0 and 11.0 showing that it has extreme alkaliphilic nature. BPT-20 and 23 also showed growth at pH 9.0, 10.0 and 11.0. Their maximum growth obtained after 22 and 5 hours of incubation respectively. Growth for BPT-06,08, and 25 was found insignificant on alkaline media of variable pH.

4.1.7. Diversity of microorganisms based on colony forms, microscopic studies, molecular characterization

Isolates obtained from NIBGE culture collection were originally isolated by Nasrin et al (2008), According to these studies all the isolates displayed different cell morphologies when viewed under the phase contrast microscope.

These cells were mostly rod shape except BPT-08. Three of the isolates BPT-08, 23, 25 had circular colony shape and the other three BPT-05, 06, 20 had irregular colony shape with raised colony elevations. BPT-06 was indicated to contain endospores. Table.3.1.3 showed the detailed characteristic of these isolates (Nasrin et al., 2008).

4.1.8. Analytical aspects of metagenomic samples

Soil samples were analyzed to study their physicochemical properties. The average temperature at the sampling sites was 25ºC. S_A was basic in nature , S_B was neutral while S_C was acidic in nature while they were tested for pH. The acidic nature of S_C sample could be due to the desulfurization of coal as this sample was collected from coal heap in NIBGE, this sample also has electrical conductivity higher (2.60 mS/ml) than S_A and S_B (1.11, 0.37 mS/ml) . Samples were also checked for the presence of different ions. S_C was dominated by high Na^+ and K^+ ion content. All the observations pertaining to these soil samples are given in Table 3.1.6. Presence of various metals is tabulated in Table 3.1.7 which indicates that percentage of Iron, Nickel and Zinc is comparatively higher than other metals.

4.2. MOLECULAR STUDIES FOR RECOVERY OF PROTEASE GENE

4.2.1. Protease gene isolation, amplification, cloning and sequencing

The study of microbial communities in natural environments has been greatly facilitated by the appliance of molecular biology tools. Isolation of DNA from bacterial cultures and direct isolation of DNA/RNA has led to better understanding of important characteristics in bacterial communities. The facts about physiology and function of bacteria has been improved by 16S rRNA analysis used as phylogenetic marker (Liesack et al.,1997).

DNA isolation from cultured bacterial isolates resulted in un sheared purified DNA (Fig3.2.1). Results of direct DNA isolation from soil samples is shown in Fig3.2.2.

There is no doubt that the investigation and applicability of functional macromolecules has led to the exploration of wide range of structural genes identified in various habitations. Yet the development of advanced molecular methods has made it possible to search and analyze for abundance and diversity of genes of interest.

Bacterial alkaline proteases are widely studied in variety of microorganisms, main source for these proteases are *bacillus* species, including as *B. licheniformis, B. subtilis, B. amyloiquefaciens, B.pumilus* and *B. alcalophilus, B. subtilis subsp. amylosacchariticus,* and the alkalophilic *Bacillus* sp. strain YaB etc. Genes from some of the above strains have been cloned and sequenced. The determination of amino acid sequence for serine proteases from *Bacillus amyloliquefaciens, Bacillus licheniformis,* and *Bacillus subtilis* has also carried out (LAAN et al.,1991 and Jiao pan et al.,2004). Most of proteases produced by these organisms are of industrial use, for the enhancement of industrial enzyme production multiple copies of genes are produced by host orgamism (Noirot et al., 1987). Present study involved the recovery of protease gene from DNA isolated from cultured bacterial isolates as well as metagenomic DNA using protease gene specific primers. Protease gene amplification for all bacterial isolates under study is presented in Fig 3.2.3 (A & B). Out of these six isolates, BPT-05 and BPT-08 showed good amplification for protease gene with primer set F1&R1 and also with C1 and C2. Multiple bands for these genes were observed which were recovered by gel extraction. Similarly, protease gene for isolate BPT- 06, BPT-20, BPT-23 and BPT-25 were amplified using B1&B2 primer set. The size of amplified gene was in the range of 950-1100bp in length for these isolates. The amplified genes were designated codes as PG05, PG06, PG08, PG20, PG23 and PG25 respectively, obtained from respective isolates.

Soil samples were processed for protease gene amplification and out of the three soil samples the amplification was observed only in $S_{C.}$ (Fig.3.2.4). Multiple bands were observed in this case which were recovered through gel extraction method. The amplified gene was designated PGSc code.

Cloning of protease gene from DNA of all six isolates and metagenomic DNA were carried out using standard laboratory protocols. These clones were identified by the presence of blue white colonies on LB plates with Ampicillin/IPTG/X-Gal. Respective white colonies from different plates were picked up and grown in LB liquid media. Plasmids were isolated and clones with protease gene as insert were selected for further processing.

Cloned protease gene and PCR fragments of the respective samples were then subjected to sequencer for sequencing. Clones were sequenced using M13 primer and PCR fragments were sequenced using protease gene specific primers for each sample. Sequencing was also carried out commercially for the confirmation of sequenced protease gene. Commercial sequencing of PCR products from all these samples gave very good results (Appendix I). But DNA sequencing of PG-06 was unsuccessful both in our lab and also commercially. So it was not processed further. This may be due to contamination in PG-06 gene sample.

Direct DNA sequencing of amplified polymerase chain reaction (PCR) products was also done in case of genes (PG05,PG06, PG08, PG20, PG2, PG25 and PGSc) . This technique has several advantages over cloning of amplified DNA products. It is faster (1 day versus 3-5 days) and in DNA samples containing sequence polymorphisms both the normal and mutated sequence can be detected in the same sequencing reaction. The major problems encountered in direct sequencing of amplified DNA have been contamination by PCR primers and deoxynucleotides, and overcoming fast template renaturation (Ausubel *et al*., 1994). Sequenced PG05 gene was 1101bp in length, sequencing of PG06 remain unsuccessful, sequenced PG08 gene was 1156 bp long, sequenced PG20gene had

819bp , PG23 gene was 950bp, sequencing of PG25 gene has resulted in 810bp while PGSc gene after sequencing was 429 basepairs.

4.3. BIOINFORMATICS STUDIES OF PROTEASE GENE

4.3.1. Sequence homology searches and nucleotide sequence analysis

To find the possible homologs for a nucleotide in nucleotide sequence databases pairwise sequence alignment is the most important step. A pairwise sequence alignment compares two protein/nucleotide sequences. BLAST is one of the most widely used programs for identification of the sequence similarities. BLAST (**B**asic **L**ocal **A**lignment **S**earch Tool) is the most extensively used local alignment tool calculates the sequence similarity quantitatively. BLAST gives an expectation value, which calculates approximately the chance such alignment can be expected.

Blast program of NCBI gives the homology matches with protease gene of our samples. Complete blast results are given in Appendix II. Main results are tabulated in Table 3.8. which shows that almost all sequences shows very good homology with other proteases from *Bacillus* strains.

PG20, PG25 and PGS_C genes were more than 90% homologous to organic solvent tolerant protease gene of *Bacillus pumilus* 115b. PG23 has 68% homology with alkaline serine protease gene of *Bacillus pumilus* TMS55. PG05 has 94% sequence homology with *Bacillus licheniformis* RG2 keratinolytic protease gene.while PG08 has 93% homology with *Bacillus licheniformis* YPIA protease gene. More than 50% of sequences homologies are indicative that query sequences are closely related to template sequences. This can also be assumed that they will code for the same protein of similar function.

Multiple sequence alignment shows the evolutionary relationship between more than two sequences, in which the conservation of nucleotides can point towards a similar functional or structural role. These results can be used for the construction of phylogenetic tree to infer ancestral origin among sequences of different families. In order to find out evolutionary relationship in protease genes under this study , multiple sequence alignment between sequenced protease gene as well as protease genes from other organism were constructed . Multiple sequence alignment between PG05,PG08,PG20,PG23,PG25 and PGS_C showed considerable sequence similarity (Appendix III). The results obtained after constructing phylogenetic tree(Fig 3.3.1 (A)) showed that have evolved from common ancestors and are homologous to each other from evolutionary point of view. Multiple sequence alignment (Fig.3.3.1 (B)) and phylogenetic tree construction between PG05,PG08,PG20,PG23,PG25 and PGS_C and protease genes from other organisms were carried out resulting in close relationship between them due to having considerable sequence similarity.

Every region of DNA has six possible open reading frames, a prokaryotic gene is represented by the longest ORF for a given region of DNA .In current study this task was performed by ORF Finder at NCBI (http://www.ncbi.nlm.nih.gov/gorf/gorf.html), the Cognitor results (Fig 3.3.2) showed that +2 and +2 open reading frame in PG05 showed significant match with subtilisin like serine protease gene, ORF -1 matches with Molybdopterin biosynthesis enzyme while other ORFs did not give any matches, indicating that these ORFs are novel. PG-08 has ORF +2 , PG20 and PG23 has ORF+1 ,PG 25 has ORF +3 matching with subtilisin like serine protease gene, while the ORFs which did not

had any matches are assumed to be novel. Pulling out of complete ORFs supports the idea that these genes are very efficient to produce and express respective proteases.

Various restriction sites are identified in protease genes under study. Fig. 3.2.2 spots out a number of restriction enzyme recognition sites in PG05, PG08,PG20,PG23,PG25 and PGS_C by building maps this information is helpful in manipulating the protease gene for further analysis in future studies.

Sequenced protease genes were then translated into respective proteins. Sequence homology searches for all translated protein sequences were carried out by using protein BLAST at NCBI. PG05 and PG08 showed 65% and 66% homology to subtilisin calsberg from *B. Subtilis* respectively. PG20 was 98% homologous to Keratinase precursor protein from *B.pumilus.* PG23 had 43% homology to serine alkaline protease preproprotein from *B.pumilus.* PG25 showed 98% homology to organic solvent tolerant protease of *B.pumilus.* where as PGSc was 82% homologus to subtilisin protein from *B.pumilus*. Prokaryotic serine proteases are discriminated by the presence of similar arrangement of catalytic His, Asp and Ser residue (Sezien .,etal 1991) there is found His active site motif (HGTHVAGTIAA) in PG05, PG08, PG20 and PG25. There is considerable active site conservation in active site region of these proteases. His, Asp and Ser catalytic triad is present in PG05, PG08, PG20 and PG25. This motif and catalytic triad is not present in PG23 and PGSc. PG23 has a Cys rich region. Protein phylogenetic results show that PG05, PG08, PG20, PG23, PG25 and PGSc have common ancestary. According to the results PG05 &PG08, PG20 & PG25, PG23 & PGSc lie close to each other having common ancestral genes with little divergence.

4.3.2. Protein modeling and in-silico analysis

On the basis of sequence similarity comparative protein models were built to analyze predicted protease structure . while comparing the 3D structures of different proteases it can be suggested that through comparative modeling functionally important regions of closely related protiens can be modelled with high precision. The tertiary structure of PG05 comprises of α / β/ α structural topology ,PG08 , PG20 and PG 25 has 2 β α β units while PG23 and PGS_C has 1 β α β unit. The overall 3D structures were visualized and studied through 3D graphical softwares. Ramachandran plot indicates that 88.4%, 87.8%, 85.4%, 83.6%, 84.7%, 91.9% of the main-chain dihedral angles are found in the most favored regions for PG05, PG08, PG20, PG23, PG25 and PGSc respectively. There are zero residues in disallowed region for all protein models except for PG23 which has 0.8% residues in the disallowed region. Prosa results shows z-score values below zero for all protein models. These results indicates that predicted models are of good quality.

Conclusion

- Current studies showed that BPT05,BPT06, BPT08,BPT20, and BPT25 were moderate halophiles, while BPT23 presented extreme halophlic characteristics. BPT05 was of extreme alkaliphilic nature while all other isolates were alkalotolerent , when tested in halophilic and alkaliphilic media for growth.
- The samples collected for protease gene isolation were of diverse nature, two different approaches were used for gene isolation, i.,e culture dependent and culture independent method.
- Pair wise sequence comparison of sequenced protease genes showed that *B.amyloliquefaciens* PG05 has 94% homology with *B.licheniformis* keratinolytic protease gene, *S.arlettae* PG08 has 93% homology with *B.licheniformis* PG20 and *B.pumilus* PG25 has 99% and 98% homology respectively with *B.pumilus* organic solvent tolorent protease gene, while *B.licheniformis* PG23 has 68% homology with *B.pumilus* alkaline serine protease gene.
- Multiple alignment and phylogentic results depicted that these genes were originated from common ancestors and they have close evolutionary relationship
- Open reading frame statistics resulted in homology of some open reading frames in the sequenced genes with subtilisin like serine proteases, most of the open reading frame in all the genes under study did not have be identified earlier, hence they were assumed to be novel.
- The present study also validates the notion that the genes of BPT05, BPT06, BPT08, BPT20, and BPT25 can play important role for the production of proteases for industry.
- Translated proteins showed that PG05 and PG08 showed 65% and 66% homology to subtilisin calsberg from *B. Subtilis* respectively. PG20 was 98% homologous to

Keratinase precursor protein from *B.pumilus.* PG23 had 43% homology to serine alkaline protease preproprotein from *B.pumilus.* PG25 showed 98% homology to organic solvent tolerant protease of *B.pumilus.* where as PGSc was 82% homologus to subtilisin protein from *B.pumilus*

- There is shown the presence of conserved His active site motif (HGTHVAGTIAA) in PG05, PG08, PG20 and PG25. There is considerable active site conservation in active site region of these proteases. His, Asp and Ser catalytic triad is present in PG05, PG08, PG20 and PG25. This motif and catalytic triad is not present in PG23 and PGSc. PG23 has a Cys rich region. High degree of conservation shows that these proteases will have same function as their template proteins.
- There is variability in structural topology of all the proteases under study.
- In future these isolated genes and comparatively modeled proteins will provide an insight to better understand stability , mechanism of action and also substrate specificity in enzyme catalysis of these proteases.
- The identification of essential core frame work in modeled proteins can use to study the active site of these proteases.
- On the basis of comparison with other proteases putative processing sites of proteases from these isolates can further expand its industrial application.
- Further these microbes could provide us with lot of different industrially significant enzymes, which have already been reported in related microorganisms.
- Similarly basic studies of the phenomenon how such microbes adopt to the high level of salt and pH provide us an interesting working model to study the mechanism how these microbes adopt to high level of salt and pH.

CHAPTER

5.0

REFRENCES

Akbalık, G. (2004), Screening for industrially important extracellular enzymes Izmir province by 16S-ITS rDNA RFLP, *J. App. Microbiol*., 766–773.

Albertini, A. M. and Galizzi, A. (1985), Amplification of a chromosomal region in Bacillus subtilis, *J. Bacteriol*., 162, 1203-11.

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

Anwar, A. and Saleemuddin, M. (1998), Role of naturally occurring autoantibodies in senescence of normal and ATP depleted goat erythrocytes, *Biochem. Mol. Biol. Int*. 44, 807-13.

Arnold, K., Bordoli, L., Kopp, J. and Schwede, T. (2006), The SWISS-MODEL Workspace: A web-based environment for protein structure homology modeling, *Bioinformatics*, 22, 195-201.

Aronson, A. I. (1993), Insecticidal toxins in *Bacillus subtilis* and other Gram-positive bacteria, *American Soc. Microbiol*, Washington D. C. 953-964.

Asmar, F., Eiland, F. and Nielsen, N. E. (1992), Inter-relationship between extracellular enzyme activity, ATP content, total counts of bacteria and CO_2 evolution, *Biology and Fertility of Soils* 14, 288–292.

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA and Struhl K (eds) (1994) *Current Protocols in Molecular Biology*. Greene Publishing Associates, Inc., and John Wiley & Sons, Inc., Boston, MA

Ayse, G., Didem, O. and Yuksel, G. (2005), Partial purification and characterization of protease enzyme from Bacillus subtilis megatherium, *Appl. Biochem. Biotechnol*, 24, 335-345.

Bach, H. J. and Munch, J. C. (2000), Identification of bacterial sources of soil peptidases, *Biology and Fertility of Soils,* 31, 219-224.

Bach, H. J., Hartmann, Schloter, A. M. and Munch, J. C. (2001), PCR primers and functional probes for amplification and detection of bacterial genes for extracellular peptidases in single strains and in soil, *Journal of Microbiological Methods,* 44, 173-182.

Badalucco, L., Kuikman, P. J. and Nannipieri, P. (1996), Protease and deaminase activities in wheat rhizosphere and their relation to bacterial and protozoan populations, *Biology and Fertility of Soils*, 23, 99-104.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. and Yeh, L. S. (2005), The Universal Protein Resource (UniProt), *Nucleic Acids Res.*, 33, D154-159.

Bakonyi, T., E. A. Gould, J. Kolodziejek, H. Weissenböck, and N. Nowotny. (2004). Complete genome analysis and molecular characterization of Usutu virus that emerged in Austria in 2001: comparison with the South African strain SAAR-1776 and other flaviviruses. *Virology* 328:301-310.

Beg, Q. K., Sahai, V. and Gupta, R. (2003), Statistical media optimization and alkaline protease production from Bacillus mojavensis in a bioreactor, *Process Biochemistry*, 39, 203.

Beynon, J. R., Prakasham, S., Subba, R. C. and Rao, S. (1989), Biotechnological aspects of microbial proteases, *Microbiol. Mol. Biol. Rev*, 62, 597-635.

Bhosale, S. H., Rao, M. B., Deshpande, V. V. and Srinivasan, M. C. (1995), Thermostability of high activity alkaline protease from *Conidiobolus coronatus* (NCL 86.8.20), *Enz. Microbial. Technol.* 17, 136-139.

Brian, A. and Fergus, P. Modern bacterial taxonomy. Van Nostrand Reinhold (U. K.), 1986; 43-72.

Brown, J. R. and Doolittle, W. F. (1997), Archaea and the prokaryote to eukaryote-transition, *Microbiol. Mol. Biol. Rev*, 61, 456-502.

Bourne, P. E., and Weissig, H. (2003), Structural Bioinformatics. John Wiley & Sons, Inc.,Hoboken, New Jersey.

Carlos Lo´ pez-Otí and Judith S. Bond, (2008). Proteases: Multifunctional Enzymes in Life and Disease, *J.biol. chem*. vol. 283, NO. 45, pp. 30433–30437.

Chang, C. N., Nielsen, J. B., Izui, K., Blobel, G. and Lampen, J. O. (1982), Identification of the signal peptidase cleavage site in Bacillus licheniformis prepenicillinase, *J. Biol. Chem*, 257, 4340-4.

Chen, S. T., Hsiao, S. C. and Wang, K. T. (1991), Stable industrial protease catalysed peptide bond formation in organic solvent, *Bioorg. Med. Chem. Letter*, 1, 445-450.

Connaris, H., West, S. M., Hough, D. W. and Danson, M. J. (1998), Cloning and overexpression in Escherichia coli of the gene encoding citrate synthase from the hyperthermophilic Archaeon Sulfolobus solfataricus, *Extremophiles*, 2, 61-66.

Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. and Barton, G. J. (1998), JPred: a consensus secondary structure prediction server, *Bioinformatics*, 14, 892-893.

Dapeau, G. R. (1976), Methods in enzymology, *Academic press, New York*, 14, 471.

Denizci, A. A., Kazan, D., Abein, E. C. A. and Eraslan, A. (2004), Newly isolated Bacillus clausii GMBAE 42, an alkaline protease producer capable to grow under high alkaline conditions, *J. App. Microbiol*, 96, 320-327.

De-Kreij, A. Venema, G. and Van den Berg, B. (2000), Substrate specificity in the highly heterogeneous M4 peptidase family is determined by a small subset of amino acids, *J. Biol. Chem,* 275, 31115-31120.

Egorova, K. and Antranikian, G. (2005), Industrial relevance of thermophilic Archaea, *Curr. Opin. Microbiol*, 8, 649-655.

Ellaiah. P., Raju, K. V., Adinarayana, K., Adinarayana, G, Prabhakar, T. and Premkumar, J. (2002), Bioactive rare actinomycetes from indigenous natural substrates of Andhra Pradesh, *Hindustan. Antibiot. Bull*, 44, 17-24.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., Pieper, U., Stuart, A. C., Marti-Renom, M. A., Madhusudhan, M. S., Yerkovich, B. and Sali, A. (2003), Tools for comparative protein structure modeling and analysis, *Nucleic Acids Res*., 31, 3375-3380.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., Pieper, U., Stuart, A. C., Marti-

Falb, M., Pfeiffer, F., Palm, P., Rodewald, K., Hickmann, V., Tittor, J. and Oesterhelt, D. (2005), Living with two extremes: conclusions from the genome sequence of Natronomonas pharaonis, *Genome. Res*. 15, 1336-43.

Ferrari, E., Jarnagin, A. S. and Schmidt, B. F. (1993), Commercial production of extracellular enzymes. In: Bacillus subtilis and other Grampositive bacteria, *American Soc. for Microbiol. Washington D. C,* 917-938.

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., Robeson, M., Edwards, R. A., Felts, B., Rayhawk, S., Knight, R., Rohwer, F. and Jackson, R. B. (2007), Metagenomic and small-subunit RNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil, *Applied and Environmental Microbiology*, 73, 21, 7059-7066.

FitzGibbon, S., Choi, A., Miller, J., Stetter, K., Simon, M., Swanson, R. and Kim, U. (1997), A fosmid-based genomic map and identification of 474 genes of the hyperthermophilic archaeon Pyrobaculum aerophilum, *Extremophiles*, 1, 36-51.

Fugishige, A., Smith, K. R., Silen, J. L. and Agard, D. A. (1992), Correct folding of alpha- lytic protease is required for its extracellular secreation from E. Coli, *J. Cell. Biol*, 118, 33-42.

Fujiwara, N. and Yamamoto, K. (1987), Production of alkaline protease in a low-cost medium by alkalophilic Bacillus sp. and properties of the enzyme, *Journal of Fermentation Technology*, 65, 345.

Fujiwara, N. and Yammato, K. (1988), Production of alkaline protease in low-cost medium by alkaliphilic Bacillus sp. and properties of enzymes. *J. Fermentation Technol*, 65, 345-350.

Fukumoto, J. (1943), Studies on the production of bacterial amylase. I. Isolation of bacteria secreting potent amylases and their distribution (in Japanese), *J. Agr. Chem. Soc. Japan*, 19, 487-503.

Germano, I., Kanzawa, T., Kondo, Y., Ito, H. and Kondo, S. (2003), Induction of autophagic cell death in malignant glioma cells by arsenic trioxide, *Cancer. Res*, 63, 2103-8.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., Bairoch A. (2003). The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acid Res.* 31(13):3784.

Gessesse, A. (1997), The use of nug-meal as a low cost substrate for the production of alkaline protease by alkaliphilic Bacillus sp. AR-009 and some properties of the enzyme, *Biores. Tech*, 62, 56-61.

Ghauri, M. A., Khalid, A. M., Grant, S., Heaphy, S. & Grant, W. D. (2003). Phylogenetic analysis of different isolates of *Sulfobacillus* spp. isolated from uranium-rich environments and recovery of genes using integron-specific primers. *Extremophiles* 7, 341–345.

Giovannoni, S. The polymerase chain reaction. *In Nucleic Acid Techniques in Bacterial Systematics*. (Eds. E Stackebrandt and M Goodfellow). 1991; 177-203. John Wiley & Sons, New York.

Greden, B. D. and Keller, M. (2006), Capturing the uncultivated majority, Applied Microbiology and Biotechnology, 17, 236–240.

Griffin, H. L., Porto, T.S. Porto and M.A. Cotta. 1992. Isolation and characterization of an alkaline protease from marine shipworm bacterium. *Curr. Microbiol*. 24: 111-117.

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

Gupta, A., Ipsita, R., Patel, R. K., Singh, S. P., Mare, S. K. and Guptam, M. N. (2005), One step purification and characterization of an alkaline protease from haloalkaliphilic Bacillus sp, *J. chromatograph*, 1075, 103-108.

Gupta, R., Beg, Q. K., Khan, S., Chauhan, B. (2002a), An overview on fermentation, downstream processing and properties of microbial alkaline proteases, *Applied Microbiology and Biotechnology*, 60, 381, (2002a).

Gupta, R., Beg, Q. K. and Lorenz, P. (2002b), Bacterial alkaline proteases: molecular approaches and industrial applications, *Applied Microbiology and Biotechnology*, 59, 15-32.

Harington, A., Watson, T. G., Louw, M. E., Rodel, J. E. and Thomson, J. A. (1988), Stability during fermentation of a recombinant ox-amylase plasmid in Bacillus subtilis. *Appl. Microbiol. Biotechnol*, 27, 521-527.

Hallberg KB, Barrie Johnson D (2001) Biodiversity of acidophilic prokaryotes. In: Advances in Applied Microbiology, Vol Volume 49. Academic Press, p 37-84

Hartshorn, M. J. (2002), AstexViewer: A visualisation aid for structure-based drug design, *J. Comput. Aided Mol. Des*, 16, 871-881.

Haulon, G. H., Hodges, N. A. and Russell, A. D. (1982), The influence of glucose, ammonium and magnesium availability on the production of protease and bacitracine by Bacillus licheniformis, *J. Gen. Microbiol*, 128, 845-851.

Hayano, K., Takeuchi, M., Ichishima, E. (1987), Characterization of a metalloproteinase component extracted from soil, *Biology and Fertility of Soils*, 4, 179-183.

Hoffmaster, A., Hill, K., Gee, J., Marston, C., De, B., Popovic, T., Sue, D., Wilkins, P., Avashia, S., Drumgoole, R., Helma, C., Ticknor, L., Okinaka, R. and Jackson, J. (2006), Characterization of *Bacillus cereus* Isolates Associated with Fatal Pneumonias: Strains Are Closely Related to *Bacillus anthracis* and Harbor *B. anthracis* Virulence, *Journal of Clinical Microbiology*, 44, 9, 3352-3360.

Horikoshi, K. (1971), Production of alkaliphilic microorganism Part I. Alkaline protease production by *Bacillus* no. 221, *Agric. Biol. Chem*, 35, 1407-1414.

Horikoshi, K. (1996), Alkaliphiles from an industrial point of view, *FEMS Microbiology Reviews*, 18, 259.

Horikoshi, K. (1999), Alkaliphiles: Some applications of their products for biotechnology, *Microbiology and Molecular Biology Reviews*, 63, 735.

Huang, Q., Peng, Y., Li, X., Wang, H. and Zhang, Y. (2003), Purification and characterization of an extracellular alkaline serine protease with dehairing function from Bacillus pumilus, *Current Microbiology*, 46, 169.

Humphrey, W., Dalke and A., Schulten, K. (1996), VMD: visual molecular dynamics. *J. Mol. Graph.*, 14, 33-38, 27-28.

Imanaka, H., Fukui, T., Atomi, H. and Imanaka, T. (2002), Gene cloning and characterization of fructose-1, 6-biphosphate aldolase from the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1, *J. Biosc. Bioeng*, 94, 237-243.

Imanaka, T., FuJii, M. and Aiba, S. (1981), Isolation and characterization of antibiotic resistance plasmids from thermophilic Bacilli and construction of deletion plasmids, *J. Bacteriol*, 146, 1091-1097.

Jacobs, M., Eliasson, M., Uhlen, M. and Flock, J. I. (1985), Cloning, sequencing and expression of subtilisin Carlsberg from Bacillus licheniformis, *Nucleic. Acids. Res*. 13, 8913-8926.

Jasvir, S., Gill, N., Devasahayam, G. and Sahoo, D. K. (1999), Studies on alkaline protease produced by Bacillus sp. NG312, *Appl. Biochem. Biotechnol*, 76, 57-63.

Johnvesly, B. and Naik, G. R. (2001), Studies on production of thermostable alkaline protease from thermophilic and alkaliphilic Bacillus sp. JB-99 in a chemically defined medium, *Process Biochemistry*, 37, 139.

Joo, H. S. and Chang, C. S. (2005), Oxidant and SDS-stable alkaline protease from a halo-tolerant Bacillus clausii I-52: enhanced production and simple purification, *J. Appl. Microbiol*, 98, 491-497.

Joo, H. S., Kumar, C. G., Park, G. C., Paik, S. R. and Chang, C. S. (2003), Oxidant and SDS-stable alkaline protease from Bacillus clausii I-52: production and some properties, *J. Appl. Microbiol*, 95, 267-72.

J. Vakhlu, Avneet Kour Sudan and B. N. Johri,(2008), Metagenomics: Future of microbial gene mining, Indian Journal of Microbiology 48, 202-215

Joo, H. S., Kumar, C. G., Park, G. C., Paik, S. R. and Chang, C. S. (2004), Bleach-resistant alkaline protease produced by a Bacillus sp. Isolated from the Korean polychaeta, Periserrula leucophryna, *Process Biochemistry*, 39, 1441-1447.

Joseph, G. and Walter, S. (2004), Extremophiles and extremozymes, Food Technol, *Biotechnol.* 42, 4, 223-235.

Juck, D., Charles, T., Whyte, L.G., Greer, C.W. (2000). Polyphasic microbial community analysis of petroleum hydrocarbon-contaminated soils from two northern Canadian communities. *FEMS Microbiol. Ecol.* 33: 241-249.

Jukes, T.H., and Cantor, C.R. Evolution of protein molecules. In: Munro, H. N (ed) *Mammalian protein metabolism*. Academic, New York. 1969; 21-132.

Kaneko, R., Koyama, N., Tsai, Y. C., Juang, R. Y., Yoda, K. and Yamasaki, M. (1989), Molecular cloning of the structural gene for alkaline elastase YaB, a new subtilisin produced by an alkalophilic *Bacillus* strain, *J. Bacteriol*, 171,5232-5236.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K. and Kikuchi, H. (1998), Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3 (supplement), *DNA. Res*, 5, 147-155.

Kobayashi, T., Hakamada, Y., Hitomi, J., Koike, K. and Ito, S. (1996), Purification of alkaline proteases from a Bacillus strain and their possible interrelationship, *Applied Microbiology and Biotechnology*, 45, 63.

Kulkarni, N., Lakshmikumaran, M. and Rao, M. (1999), Xylanase II from an alkaliphilic thermophilic Bacillus with a distinctly different structure from other xylanases: evolutionary relationship to alkaliphilic xylanases, *Biochem. Biophys. Res. Commun*, 263, 640-645.

Kumar, C. G. and Takagi, H. (1999), Microbial alkaline proteases: from a bioindustrial Viewpoint, *Biotechnology Advances*, 17, 561.

Ladd, J. N., Butler, J. H. A. (1972), Short-term assays of soil proteolytic enzyme activities using proteins and dipeptide derivatives as substrates, *Soil Biology & Biochemistry*, 4, 19-30.

Lantz, M.S. and Cibrowski, P. (1994) Zymographic techniques for detection and characterization of microbial proteases. *Methods Enzymol* **235**: 563–594.

Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. and Thornton, J. M. (1996), AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR, *J. Biomol. NMR.*, 4, 477-486.

Layman, P. L. (1986), Industrial enzymes: batteling to remain specialist, *Chem. Eng. News*, 64, 11-14.

MacElroy, R. D. (1974), Some comments on the evolution of extremophiles, 6,74-75.

Madigan, M. T. and Marrs, B. L. (1997), Extremophiles, *Sci. Am*, 276, 82-87.

Manachini, P. L., Rosebrough, N. J., Forr, A. L. and Randall, J. R. (1988), Thermostable alkaline protease produced by Bacillus thermoruber, a new species of Bacillus, *Biotechnol*, 28, 409-413.

Markland, F. S. and Smith, E. L. (1971), Subtilisins: primary structure, chemical and physical properties, *Academic. Press. Inc, NewYork*, 3, 561-608.

McGinnis, Scott. and Madden, L. T. (2004), BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, 32, W20-25.

McIver, K. S. and Scott, J. R. (1997), Role of mga in growth phase regulation of virulence genes of the group A streptococcus, *J. Bacteriol*, 179, 5178-5187.

Mongodin, E. F., Nelson, K. E., Daugherty, S., Deboy, R. T., Wister, J., Khouri, H., Weidman, J., Walsh, D. A., Papke, R. T., Sanchez Perez, G., Sharma, A. K., Nesbo, C. L., MacLeod, D., Bapteste, E., Doolittle, W. E., Charlebois, R. L., Legault, B. and Montville, J. T. (1983), Dual-substrate plate diffusion assay for proteases, *Appl. Env. Microbiol*, 45, 200-204.

Nakanishi, T. and Yamamoto, T. (1974), Action and specificity of a *Streptomyces* alkaliphilic proteinase, *Agric. Biol. Chem*, 38, 2391-2397.

Nannipieri, P., Muccini, L. and Ciardi, C., (1983), Microbial biomass and enzyme activities: production and persistence, *Soil Biology & Biochemistry*, 15, 679-685.

Nathalie, W., Ursula, G., Ajay, S., Julian, B. C., David, W. H. and Michael, J. D. (2003), Use of a Packed-Column Bioreactor for Isolation of Diverse Protease-Producing Bacteria from Antarctic Soil, *Applied And Environmental Microbiology*, 1457-1464.

Nedkov, P., Oberthur, W. and Braunitzer, G. (1985), Determination of the complete amino-acid sequence of subtilisin DY and its comparison with the primary structures of the subtilisins BPN, Carlsberg and amylosacchariticus, *Biol. Chem. Hoppe. Seyler*, 366, 421-30.

Nielsen, P., Fritze, D. and Priest, G. (1995), Phenitic diversity of alkaliphilic Bacillus strains: proposal for nine new species, *Microbiology*, 141, 1745-1761.

Noirot, P., M. A. Petit and S. D. Ehrlich. (1987). Plasmid replication stimulates DNA recombination in Bacillus subtilis. *J. Mol. Biol.* 196:39–48.

Okuta, A., Ohnishi, K. and Harayama, S. (1998), PCR isolation of catechol 2, 3-dioxygenase gene fragments from environmental samples and their assembly into functional genes, *Gene*, 212, 221-228.

Olsen, G. J., Lane, Giovannioni, S. J., and Pace, N. R. Microbial ecology and evolution: a Ribosomal RNA approach. Ann. *Rev Microbial*. 1986; **40**, 337-365.

Oren, A., Heldal, M., Norland, S. and Galinski, E. A. (2002), Intracellular ion and organic solute concentrations of the extremely halophilic bacterium Salinibacter ruber, *Extremophiles*, 6, 491-498.

Ouoba, L. I., Cantor, M. D., Diawara, B., Traore, A. S. and Jakobson, M. (2005), Degredation of African locust bean oil by Bacillus subtilis and Bacillus Pumilis isolated from soumbala, a fermented African locust bean condiment, *J. Appl. Microbiol*, 95, 4, 868-873.

Palva, I., Pettersson, R. F., Kalkkinen, N., Lehtovaara, P., Sarvas, M., Soderlund, H., Takkinen, K. and Kaariainen, L. (1981), Nucleotide sequence of the promoter and NH_2-terminal signal peptide region of the alpha-amylase gene from Bacillus amyloliquefaciens, *Gene*. 15, 43-51.

Park, Y.S., J.K. Lee, Y.O. Kim, H.K. Kim and T.K. Oh. (1996). Purification and characterization of a thermostable alkaline protease from *Thermoactinomyces sp*. E79 and the DNA sequence of the encoding gene. *Biosci. Biotechnol. Biochem*. 60:840-6.

Pazel, D. P. (1989), DS-Viewer- interactive graphical data structure presentation facility, *IBM Systems J.*, 28, 307-323.

Polgar, L. (1989), Mechanisms of protease action. CRC Press, Inc., Boca Raton, Florida.

Powers, S. D., Adams, R. M. and Wells, J. A. (1986), Secretion and autoproteolytic maturation of subtilisin, *Proc. Natl. Acad.Sci, USA*, 83, 3096-3100.

Prakasham, R. S., Rao, S., Rajesham, S. and Sarma, P. N. (2005), Optimization of alkaline protease production by Bacillus sp using Taguchi methodology, *Appl. Biochem. Biotechnol*, 120,133-144.

Priest, F. G. (1977), Extracellular enzyme synthesis in the genus Bacillus, *Bacteriol. Rev*, 41, 711-753.

Priest, F. G., Goodfellow, M., Shute, L. A. and Berkeley, R. C. W. (1987), Bacillus amyloliquefaciens sp. nov. nom. rev. Int, J. Syst. Bacteriol, 37, 69-71.

Rahman, R. N. Z. A., Razak, C. N., Ampon, K., Basri, M., Yunus, W. M. Z. W. and Salleh, A. B. (1994), Purification and characterization of a heat-stable alkaline protease from Bacillus stearothermophilus F1, *Applied Microbiology and Biotechnology*, 40, 822.

Rao, M. B., Tanksale, A. M., Mohini, S. G. and Deshpande, V. V. (1998), Molecular and Bio-technological aspects of microbial proteases, *Microbiology and Molecular Biology Reviews*, 62, 597.

Renom, M. A., Madhusudhan, M. S., Yerkovich, B. and Sali, A. (2003), Tools for comparative protein structure modeling and analysis, *Nucleic Acids Res*., 31, 3375-3380.

Rodriguez-Valera, F. (2005), The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea, *Proc. Natl. Acad. Sci. USA*, 102, 18147-18152.

Rost, B. (1996), PHD: predicting one-dimensional protein structure by profile-based neural networks, *Methods Enzymol*, 266, 525-39.

Rothschild, L. J. and Mancinelli, R. L. (2001), Life in extreme environments, *Nature*, 409, 1092-1101.

Ruepp, A., Graml, W., Santos-Martinez, M. L., Koretke, K. K., Volker, C., Mewes, H. W., Frishman, D., Stocker, S., Lupas, A. N. and Baumeister, W. (2000), The genome sequence of the thermoacidophilic scavenger Thermoplasma acidophilum, *Nature*, 407, 508-513.

Saeki, K., Hitomi, J., Okuda, M., Hatada, Y., Kageyama, Y., Takaiwa, M., Kubota, H., Hagihara, H., Kobayashi, T., Kawai, S. and Ito, S. (2002), A novel species of alkaliphilic Bacillus that produces an oxidatively stable alkaline serine protease, *Extremophiles*, 6, 65-72.

Saitou, N., and Nei, M. The neighbour–joining method: a new method for reconstruction phylogenetic trees. *Mol Biol Evolu*. 1987; **4**, 406-425.

Sakurai, M., Suzuki, K., Onodera, M., Shinano, T. and Osaki, M. (2007), Analysis of bacterial communities in soil by PCR–DGGE targeting protease genes, *Soil biology & biochemistry*, 39, 2777-2784.

Silvanovich, A., Bannon, G. and McClain, S. (2009), The use of E-scores to determine the quality of protein alignments, *Regul Toxicol Pharmacol*, 54, 26-31.

Sippl, M. J, (1993), Recognition of errors in three-dimensional structures of proteins, *Proteins*, 17, 355-362.

Salkinoja-Salonen, S., Vuorio, R., Andersson, M. A., Kämpfer, P., Andersson, M. C., Honkanen-Buzalski, T. and Scoging, A. C. (1999), Toxigenic Strains of Bacillus licheniformis Related to Food Poisoning, *Appl Environ Microbiol*, 65, 10, 4637-4645.

Sanchez, P. C., Mellado, E., Mertoldo, C., Antranikian, G. and Ventosa, A. (2003), Screening and characterization of protease CP1 produced moderately halophilic bacterium of Pseudomonas sp. Strain CP-76, *Extremophiles*, 7, 221-228.

Schallmey, M., Singh, A. and Ward, O. P. (2004), Developments in the use of Bacillus species for industrial production, *Can. J. Microbiol*, 50, 1-16.

Schleifer, K. H., Kilpper-Balz, R. and Devriese, L. A.(1985), Staphylococcus arlettae sp. nov. S. equorum sp. nov. and S. kloosii sp. Nov, three new coagulase-negative, novobiocin-resistant species from animals. Syst, *Appl. Microbiol*, 5, 501-509.

Shibu, Yooseph1., Granger, Sutton., Douglas, B., Rusch, A., Halpern, L., Shannon, J. W., Karin, R., Jonathan, A. E., Karla, B. H., Gerard, M., Weizhong, L., Lukasz, J., Piotr, C., Christopher, S. M., Huiying, L., Susan, T. M, Marcin, P. J, Christopher, V. B, John-Marc, C., David, A. S., Yufeng, Z., Kannan, N., Shaun, L., Benjamin, J. R., Vineet, B., Robert, F1., Steven, E. B., Adam, G., David, E., Jack, E. D., Susan, S. T., Robert, L. S., Marvin, F. and Venter, J. C. (2007), The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families, *PLoS Biology*, 0432-0466.

Shumi, W., Towhid, M. D. and Anwar, M. N. (2004), Proteolytic activity of bacterial isolate Bacillus fastidiosus den dooren de jong, *Biol. Sci*, 4, 3, 370-374.

Silvanovich, A., Bannon, G. and McClain, S. (2009), The use of E-scores to determine the quality of protein alignments, *Regul Toxicol Pharmacol*, 54, 26-31.

Sippl, M. J, (1993), Recognition of errors in three-dimensional structures of proteins, *Proteins*, 17, 355-362.

Strauch, M. A. and Hach, J. A. (1993), Transition-state regulators: sentinels of Bacillus subtilis post exponential phase gene expression, *Mol. Microbiol*, 7, 337-342.

Studdert, C. A., Herrera, M. K., Seitz, M. I., Plasencia, G., Sanchez, J. J. and Castro, R. E. (2001), Purification and biochemical characterization of the haloalkaliphilic archaeon Natronococcus occultus extracellular serine protease, *J. Basic. Microbiol*, 41, 375-383.

Takami, H. and Horikoshi, K. (2000), Analysis of the genome of an alkaliphilic Bacillus strain from an industrial point of view, *Extremophiles*, 4, 99-108.

Takami, H. and Krulwich, T. A. (2000), Reidentification of facultatively alkaliphilic Bacillus firmus OF4 as Bacillus pseudofirmus OF4, *Extremophiles*, 4, 19-22.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997), The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res.*, 25, 4876-4882.

Thorne, C. B. (1993), Bacillus anthracis. In: Bacillus subtilis and other Gram-positive bacteria, *American Society for Microbiology Washington D.C.* 113-124.

Torsvik, V., Goksoyr, J. and Daae, F. L. (1990), High diversity in DNA of soil bacteria, *Applied and Environmental Microbiology*, 56, 782-787.

Valenzuela, L., Chi, A., Beard, S., Orell, A., Guiliani, N., Shabanowitz, J., Hunt, D. and Jerez, C. (2006), Genomics, metagenomics and proteomics in biomining microorganisms, *Biotechnol. Adv*, 24, 197-211.

Van de peer, Y., and De Wachter, R. (1993), TREECON: a software package for the construction and drawing of evolutionary trees. *Comput Appl Biosci*.; **9**, 177-182.

Van de Loosdrecht, A. A., Heuff, G., Betjes, M. G., Beelen, R. H. and Meijer, S. (1995), Isolation and purification of large quantities of fresh human Kupffer cells, which are cytotoxic against colon carcinoma, *Hepatology*, 21, 740-745.

Van den Berg, B. (2003), Extremophiles as a source for novel enzymes, *Curr. Opin. Microbiol*, 6, 213-218.

Van Loosdrecht, M. C. M., Lyklema, J., Norde, W. and Zehnder, A. J. B. (1990), Influence of interfaces on microbial activity, Microbiol. Rev, 54, 75-87.

Ventosa, A., Nieto, J. J. and Oren, A. (1998), Biology of moderately halophilic aerobic bacteria, *Microbiol. Mol. Biol. Rev*, 62, 504-544.

Vermehlo, A. B., Meirelles, M. N. L., Loops, A., Petinate, S. D. G., Chaia, A. A. and Branquinha, M. H. (1996), Detection of extraellular protease from microorganisms on agar plates, *Raio de Joueiro*, 91, 6, 755-760.

Vilain, S., Luo, Y., Hildreth, M. and Brozel, V. (2006), Analysis of the Life Cycle of the Soil Saprophyte Bacillus cereus in Liquid Soil Extract and in Soil, *Applied Environmental Microbiology*, 72, 7, 4970-4977.

Vriend, G. (1990), WHAT IF: A molecular modeling and drug design program *J. Mol. Graph*., 8, 52-56.

Ward, M. W., Bateson, M. M., Weller, R., and Ruff-Roberts, A. L. (1992). Ribosomal RNA analysis of microorganisms as they occur in nature. *In Recent Advances in Microbial ecology*. Ed. K. C Marshall; 219-286. Plenum Press, New York.

Ward, O. P. (1985), Proteolytic enzymes. In: Comprehensive Biotechnology (Moo Young ed.), *Pergamon press, New York, 784*.

Watanabe, K. and Hayano, K., (1994), Estimate of the source of soil protease in upland fields, *Biology and Fertility of Soils*, 18, 341-346.

Wiederstein, M. and Sippl, M. J. (2007), ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins, *Nucleic Acids Res*., 35, W407-10.

Winson, M. K. and Kell, D. B. (1997), If you've got it, flaunt it-rapid screening for microbial biocatalysts, *Trends. Biotechnol*, 15, 120-122.

Woese, C. R., Stackebrandt, E., Macke, T. and Fox, G. E. A definition of the eubacterial "divisions". *Appl. Microbial*. 1985; **6,** 143-151.

Wong, S., Price, C., Goldfarb, D. and Doi, R. (1984), The subtilisin E gene of Bacillus subtilis is transcribed from a sigma 37 promoter in vivo, *Proc. Natl. Acad. Sci. USA*, 81, 4, 1184-1188.

Yang, J. K., Shuh, I. L. and Wsng, Y. M. (2000), Production and purification of potease from a Bacillus subtilus that can deproteinize crustacean wastes, Enzyme Microbial Technol, 26, 406-413.

Yeates, C., Gillings, M. R., Davison, A. D., Altavilla, N. and Veal, D. A. (1997), PCR amplification of crude microbial DNA extracted from soil, *Letters in Applied Microbiology.*

Young, M. and Ehrlich, S. D. (1989), Stability of reiterated sequences in the Bacillus subtilis chromosome, *J. Bacteriol,* 171, 2653-2656.

Yumoto, I., Yamazaki, K., Sawabe, T., Nakano, K., Kawasaki, K., Ezura, Y. and Shinano, H. (1998), Bacillus horti sp. nov. a new gram-negative alkaliphilic bacillus. *Int, J. Syst. Bacteriol,* 48, 2, 565-571.

Zhou, J., Bruns, M. A. and Tiedje, J. M. (1996), DNA recovery from soils of diverse composition, *Applied and Environmental Microbiology,* 62, 2, 316-322.

Zuidweg, M. H., Bos, C. J. and Welzen, H. (1972), Proteolytic components of alkaline proteases of Bacillus strains, Zymograms and electrophoretic isolation, *Biotechnol. Bioeng*, 14, 685-714.

APPENDICES

Appendix I
16S r DNA genes sequences

>3B

GGTAGCTGTNGNGTGCGNTCGAGCGGATGAAGGGAGCTTGCTCCTGGATTCAAGCGG
CGGACGGGTGAGTAATGCCTAGGAATCTGCCTGGTAGTGGGGGATAACGTCCGGAAA
CGGGCGCTAATACCGCATACGTCCTGAGGGAGAAAGTGGGGGATCTTCGGACCTCAC
GCTATCAGATGAGCCTAGGTCGGATTAGCTAGTTGGTGGGGTAAAGGCCTACCAAGG
CGACGATCCGTAACTGGTCTGAGAGGATGATCAGTCAAACTGGAACTGAGACAGGGT
CCACACTTATACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGAT
CCAGCCATGCCGCGTGTGTGAAGAAGGTCTTCGGATTGTAAAGCACTTTAAGTTGGGA
GGAAGGGCAGTAAGTTAATACCTTGCTGTTTTGACGTTACCAACAGAATAAGCACCGG
CTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGTGCAAGCGTTAATCGGAATTA
CTGGGCGTAAAGCGCGCGTAGGTGGTTCAGCAAGTTGGATGTGAAATCCCCGGGCTC
AACCTGGGAACTGCATCCAAAACTACTGAGCTAGAGTACGGTAGAGGGTGGTGGAAT
TTCCTGTGTAGCGGTGAAATGCGTAGATATAGGAAAGGAACACCAGTGGCGAAGGCG
ACCACCTGGACTGATACTGACACTGAGGTGCGAAAGCGTGGGGAGCAAACAGGATTA
GATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTAGCCGTTGGGATCCTTGAGAT
CTTAGTGGCGCACTTAACGCGT

>2A

GTGCAGTCGAGCGGATGAAGGGAGCTTGCTCCTGGATTCAGCGGCGGCACGGGTGAG
TAATGCCTAGGAATCTGCCTGGTAGTGGGGGATAACGTCCGGAAACGGGCGCTAATA
CCGCATACGTCCTGAGGGAGAAAGTGGGGGATCTTCGGACCTCACGCTATCAGATGA
GCCTAGGTCGGATTAGCTAGTTGGTGGGGTAAAGGCCTACCAAGGCGACGATCCGTA
ACTGGTCTGAGAGGATGATCAGTCACACTGGAACTGAGACACGGTCCAGACTCCTAC
GGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGCCG
CGTGTGTGAAGAAGGTCTTCGGATTGTAAAGCACTTTAAGTTGGGAGGAAGGGCAGT
AAGTTAATACCTTGCTGTTTTGACGTTACCAACAGAATAAGCACCGGCTAACTTCGTG
CCAGCAGCCGCGGTAATACGAAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAA
GCGCGCGTAGGTGGTTCAGCAAGTTGGATGTGATATCTTCGGGCTCAACCTGGGAACT
GCATCCAACACTACTGAGCTAGAGTACGGTAGAGGGTGGTGGAATTTCCTGTGTAGCG
GTGAAATGCGTAGATATCGGAAGGAGCACGAGTGGCGAAGGCGACCACCTGGGCTGA
TACTGACACTGAGGTGCGAAACGTGGGAGCAAACAGGATTAGATACCCTGGTAGTCC
ACGCCGTATACGATGTCTACTAGCTCTGGGATCTTGAGATCTTAGTCGCGCTCTAACG
CGATAAGT

>3A

GTGCAGTCGAGCGGATGACGGGAGCTTGCTCCTTGATTCAAGCGGCGGACGGGNGAG
TAATGCCTAGGAATCTGCCTGGTAGTGGGGGACAACGTTTCGAAAGGAACGCTAATA
CCGCATACGTCCTACGGGAGAAAGCAGGGGACCTTCGGGCCTTGCGCTATCAGATGA
GCCTAGGTCGGATTAGCTAGTTGGTGGGGTAATGGCTCACCAAGGCGACGATCCGTA
ACTGGTCTGAGAGGATGATCAGTCACACTGGAACTGAGACACGGTCCAGACTCCTAC
GGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGCCG
CGTGTGTGAAGAAGGTCTTCGGATTGTAAAGCACTTTAAGTTGGGAGGAAGGGCAGT
AAGCTAATACCTTGCTGTTTTGACGTTACCGACAGAATAAGCACCGGCTAACTCTGTG

CCAGCAGCCGCGGTAATACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAA
GCGCGCGTAGGTGGTTTGTTAAGTTGGATGTGAAAGCCCCGGGCTCAACCTGGGAACT
GCATCCCAAACTGGCAAGCTAGAGTACGGTAGAGGGTGGTGGAATTTCCTGTGTAGC
GGTGAAATGCGTAGATATCGGAAGGAACACCAGTGGCGAAGGCGACCACCTGGACTG
ATACTGACACTGAGGTGCGAAAGCGTGGGGAGCAAACCGGATTAGATACCCTGGTAG
TCCACGCCGTAACGATGTCAACTAGCCGTTGGAATCCTTGAGAATTTAGTGCCGCAGC
TAACGCATTAAGTTGCC

S1

TGCAGTCGAGCGGATGAAGGGAGCTTGCTCCTGGATTCAGCGGCGGCACGGGTGAGT
AATGCCTAGGAATCTGCCTGGTAGTGGGGGATAACGTCCGGAAACGGGCGCTAATAC
CGCATACGTCCTGAGGGAGAAAGTGGGGGATCTTCGGACCTCACGCTATCAGATGAG
CCTAGGTCGGATTAGCTAGTTGGTGGGGTAAAGGCCTACCAAGGCGACGATCCGTAA
CTGGTCTGAGAGGATGATCAGTCACACTGGAACTGAGACACGGTCCAGACTCCTACG
GGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGCCGC
GTGTGTGAAGAAGGTCTTCGGATTGTAAAGCACTTTAAGTTGGGAGGAAGGGCAGTA
AGTTAATACCTTGCTGTTTTGACGTTACCAACAGAATAAGCACCGGCTAACTTCGTGC
CAGCAGCCGCGGTAATACGAAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAG
CGCGCGTAGGTGGTTCAGCAAGTTGGATGTGAAATCCCCGGGCTCAACCTGGGAACT
GCATCCAAAACTACTGAGCTAGAGTACGGTAGAGGGTGGTGGAATTTCCTGTGTAGC
GGTGAAATGCGTAGATATAGGAAGGAACACGAGTGGCGAAGGCGACCACCTGTACTG
GTACTGACAGTGATGTGCGAAAGCGGGGGAGCAAACAGGATTAGATACCCTGGTAGT
CGACGCCGTATACGATGTCGACTAGCCTCTGGGATCCTTGAGAACTTAGTGGCGCACT
AACGCGTAA

S2

TGCAGTCGAGCGGGGGAAGGTAGCTTGCTACCGGACCTAGCGGCGGACGGGTGAGTA
ATGCTTAGGAATCTGCCTATTAGTGGGGGACAACATCTCGAAAGGGATGCTAATACCG
CATACGTCCTACGGGAGAAAGCAGGGGATCTTCGGACCTTGCGCTAATAGATGAGCC
TAAGTCGGATTAGCTAGTTGGTGGGGTAAAGGCCTACCAAGGCGACGATCTGTAGCG
GGTCTGAGAGGATGATCCGCCACACTGGGACTGAGACACGGCCCAGACTCCTACG
GGAGGCAGCAGTGGGGAATATTGGACAATGGGGGGAACCCTGATCCAGCCATGCCGC
GTGTGTGAAGAAGGCCTTATGGTTGTAAAGCACTTTAAGCGAGGAGGAGGCTACTCT
AGTTAATACCTAGGGATAGTGGACGTTACTCGCAGAATAAGCACCGGCTAACTCTGTG
CCAGCAGCCGCGGTAATACAGAGGGTGCGAGCGTTAATCGGATTTACTGGGCGTAAA
GCGTGCGTAGGCGGGTTATTAAGTCGGATGTGAAATCTCCGAGCTTAACTTGGGAATT
GCATTCGATACTGGTGAGCTAGAGTATGGGAGAGGATGGTAGAATTCCAGGTGTAGC
GGTGAAATGCGTAGAGATCTGGAGGAATACCGATGGCGAAGGCAGCATCTGGCTAAT
ACTGAGGCTGAGGTACGAAAGATGGGGAGCAACAGAATATATACTCTGGTAGTCCAT
GCCGTAAATATGTCTACTATTCGTTGGGCCTTTGAGGCTTTAGTGGCCGCTCTCACACT
ATAGCTA

Sequenced protease genes

>PG-05

AGTCTGGAGCTACCTCAACGTAGCCTCTCAACTCACTAGTCCTGTCGATGCAAAGAAA
TGGGCGTCATTGTCCGGAGTTTTCCATGGATTTGCAACGCGTCCATTAAGGATAAGGA
GAGCAGCAATGTCGAAGCTGCCGCTCAACTCGAGATTCCAGTGATGCGGCGGCTCCG
GCGGCGTGCCCGGATGAAAAGTCGAGGCTGCGGCGGCCTCAAAAATGAGGGGGAAC
AATGTCTGATCTGAGGCTCCCCGAAAGAATCAATCAGGCTCAAGGCCCAAGCTGCCTT
CCATGTTACGATTCAAGGGGCCTCCTTGCTGGGGAACCTGATTTAAGATGCCTCGCAC
CTCGACAAAAATGCAGATCAAATCAAGATTCATGGATACAGGAATCCAAGCTTCTCAT
CCGGACTTGAACGTATTCGGCGGACCAAGCTTTGTGGCTGGGGAAGCTTATAACACCG
ACGGCAACGGACACGGCACACATGTTGCCGGTACAGTAGCTGCGCTTGACAATACAA
CGGGTGTATTAGGCGTTGCGCCAAGCGTATCCTTGTACGCGGTTAAAGTACTGAATTC
AAGCGGAAGCGGATCATACAGCGGCATTGTAAGCGGTATCCAGTGGGCGACAACAAA
CGGCATGGATGTTATCAATATGAGCCTTGGGGGAGCATCACGCTCAACAGCGATGAA
ACAGGCAGTCGACAATGCATATGCAAGAGGGGTTGTCGTTGTAGCTGCAGCAGGGAA
CAGCGGATCTTCCGGAAAACACGAATACAATTGGCTATCCTGCAAAATACGATTCTGT
CATCGCAATTGGTGCGGTATACTCTAACAGCAACAGAGCTTCATTTTCCAGTGTGGGA
GCAGAGCTTGAAGTCATGGCTCCTGGCGCAGGTGTATACAGCTCTTACCCCACCCAAC
ACTTATGCAACATTGCACGGAACGTCAATGGCATTCTCCTCATGTAGCCGGGAGCACA
GCTATGATCTTGTCAAAACATCCCAATCTTTCAGCTTCACGAGTCCGACAAGCGTCTCT
CACACGGCTACTTATTTGCAAGCTCCTGCTACATAATCGAAAAGGTACTGATCAATGT
CAAGC

>PG-08

CGCCCCTGATGCTACGTCGCACGTAGCCGCTCATCTAACTATCCTGTCGATGCAAAGA
AATGGGCGTCATTGTCCGGAGTTTTCCATGGATTTGCAACGCGTCCATTAAGGATAAG
GAGAGCAGCAATGTCGAAGCTGCCGCTCAACTCGAGATTCCAGTGATTCCTGGGTGG
GACGGCGTGATGCTCGCCAAGGACAAAGCTGCGGCTTAACTCAAAATCCATGAAAAA
AACCATCTGGAAAAAGGCTCCCTGAAACAACAAAAAAGGACCGTGGCCAAATTGCCC
CTGAGATTCAGCCCAACGGGCTTCCTCGCATTAAAGCGGAGAAAAAGCAGGCTCTCG
GCTTTAAAGGAACGAATGTAAAAGTAGCCGTCCTGGATACAGGAATCCAAGCTTCTC
ATCCGGACTTGAACGTAGTCGGCGGAGCAAGCTTTGTGGCTGGCGAAGCTTATAACAC
CGACGGCAACGGACACGGCACACATGTTGCCGGTACAGTAGCTGCGCTTGACAATAC
AACGGGTGTATTAGGCGTTGCGCCAAGCGTATCCTTGTACGCGGTTAAAGTACTGAAT
TCAAGCGGAAGCGGATCATACAGCGGCATTGTAAGCGGTATCGAGTGGGCGACAACA
AACGGCATGGATGTTATCAATATGAGCCTTGGGGGAGCATCAGGCTCGACAGCGATG
AAACAGGCAGTCGACAATGCATATGCAAGAGGGGTTGTCGTTGTAGCTGCAGCAGGG
AACAGCGGATCTTCAGGAAAACACGAATACAATTGGCTATCCTGCGAAATACGATTCT
GTCATCGCAGTTGGGTGCGGTAGACTCTAACAGCAACAGAGCTTCATTTTCCAGTGGT
GGGGAGCAGAGCTTGAAGTCATGGCTCCTGGGCGCAAGGCGTATACAGCACTTACCA
ACCGAACACTTATGCAACATTGGAACGGAACCGTCAATGGCTTCCTCTCATGTAGCGG
AAGCAGCAGCTTTTGATCTTTGTCAAAAACATCCGAATCTTTCAGCTTCACCAAGTTCC
GGCAAACGGTCTTCTTTCAGCAACGAGCGGACCTTATTTGGGAAGGCTCCGTTCTACT
AAGGAAAAGGTCCTGAATCGATTGTCGAAAGCTGGCCCGACTCAAACATCCGAGAAA
TTTCAA

>PG-20

GGGCGCACATCTACTAGCATCCGCTCCCAACAATGCGATCAAGGTGCTAATGTCAAAA
TAGCCGTCCTTGATACTGGAATCCACGCTGCACACCCTGACTTAAATGTTGCAGGCGG
TGCGAGCTTCGTCCCTTCAGAGCCAAATGCCACCCAAGACTTTCAATCACATGGAACT
CACGTAGCTGGAACCATTGCTGCCCTTGATAACACAATTGGTGTTCTTGGGGTTGCTC
CAAACGCTTCCCTATATGCTGTGAAAGTATTAGACCGTAATGGTGACGGACAATACAG
CTGGATTATTAGCGGTATTGAATGGGCTGTAGCTAATAATATGGATGTCATCAATATG
AGCTTAGGCGGACCAAGCGGTTCAACAGCGCTTAAAAATGCCGTGGATACAGCGAAT
AACCGTGGAGTCGTTGTTGTGGCGGCCGCAGGTAATTCTGGCTCTAGTGGCTCTAGAA
GTACAGTTGGCTATCCAGCAAAATACGATTCTACAATTGCCGTTGCCAATGTAAACAG
TAACAATGTCAGAAACTCATCTTCTAGCGCAGGTCCTGAATTAGATGTTTCTGCACCT
GGTACTTCTATTTTAAGTACAGTCCCAAGCAGCGGATACACATCTTATACTGGAACAT
CTATGGCGTCTCCTCATGTAGCAGGAGCAGCAGCGCTTATTCTTTCTAAAAACCCGAA
TCTAACAAATTCACAGGTTCGCCAGCGCTTAGAAAATACAGCAACACCGCTTGGTGAC
TCATTCTATTACGGAAAAGGGTTAATCAACGTTCAAGCAGCTTCTAACTATAAAAGGA
TCCGGAAA

>**PG-23**

GNNGNACCTANNTACTTCTCCCTAATGTTCCGCTCGGGTCCTAAAGGCGTTAATGGCA
AAGTACTGGTTGTTGATTTTGCCTATAATGCTGCACATCCTGGCTTGAATGATGCATGG
GCTGCGAGGTTCCCTGCGTGAAATTGAAATGCCCTCCCAAATTTTCACTCCCTTGGAA
CTCGAATAATTGGAAGCATTGCTGGCCTTGCCTACTGAGTGGGCGTTCTTGGGGAATT
GCGTTGGCTTCCCTGCCCGCTGTCCAAGCATTAAACCTGTCTGGCCACCTGCATTAATG
ATGGATTATTAGCGCGATTGAAGGGGGTGTGCGCAATGAGAGGTATGTCGTTTCTATG
ATTTTTGACTCAGTGCGCTCGTCCGCACGGCTGAAAAAAGCGGTTTATACTCACTAAA
AGGGCGGAATACTTGTATCTGCCGAATCAGGGAATTACGGCTCAAAGAACTCTAGAA
CAAAAGGTGGCTATCCAGCCAAAAACCAGTCTAAGGTTGCTGTGCTGGATGTTTTCCG
TAACCTCGTCCCCCCTGAATCTTCTACCGAAATCCCTGATTTAGATGTTTCTGCCCCTG
CCACTTCTATTTTTAATAATACGCCGAGTTTCGCACACACTTCTCATACGGGAACTCCT
CTGGTCTCACCCTGTCGCTTACGAGAAACCTGTCTTACTTTTTCTAATAACCAAAACCT
AACCTATCTCCAAGCTCACGAGCGCTGTATAAAATACTCCGACACCGCTTGGTGACTC
ATTCTATTATGTGAAAACGAACCCCCCCGTTCACCCGACCGCTAACTAATGATCCGAA
ACTATATGCTTGCGCGAGGTACCGAGCTCACTTCTCATCGCCACTTTTTACACCCTCTT
GACTGGGAAAACCTGACGTTACCCAACTTAATCTGCTAGCGACATATCCACCTTTCGC
CTCTCGACTATACCTAG

>**PG-25**

TTTAAGTTTCTTTTGGAGGGGCTAAAAGTTTTAAGGGTGCAAATGTCAAAGTAGCCGT
CCTTGATACTGGAATCCACGCTGCACACCCTGACTTAAATGTTGCAGGCGGTGCGAGC
TTCGTCCCTTCAGAGCCAAATGCCACCCAAGACTTTCAATCACATGGAACTCACGTAG
CTGGAACCATTGCTGCCCTTGATAACACAATTGGTGTTCTTGGGGTTGCTCCAAACGC
TTCCCTATATGCTGTGAAAGTATTAGACCGTAATGGTGACGGACAATACAGCTGGATT
ATTAGCGGTATTGAATGGGCTGTAGCTAATAATATGGATGTCATCAATATGAGCTTAG
GCGGACCAAGCGGTTCAACAGCGCTTAAAAATGCCGTGGATACAGCGAATAACCGTG
GAGTCGTTGTTGTGGCGGCCGCAGGTAATTCTGGCTCTAGTGGCTCTAGAAGTACAGT
TGGCTATCCAGCAAAATACGATTCTACAATTGCCGTTGCCAATGTAAACAGTAACAAT

GTCAGAAACTCATCTTCTAGCGCAGGTCCTGAATTAGATGTTTCTGCACCTGGTACTTC
TATTTTAAGTACAGTCCCAAGCAGCGGATACACATCTTATACTGGAACATCTATGGCG
TCTCCTCATGTAGCAGGAGCAGCAGCGCTTATTCTTTCTAAAAACCCGAATCTAACAA
ATTCACAGGTTCGCCAGCGCTTAGAAAATACAGCAACACCGCTTGGTGACTCATTCTA
TTACGGAAAAGGGTTAATCAACGTTCAAGCAGCTTCTACAAAAGGAATCCCCGAAA

>PG-S_C

TTCAAACGGTAGTTATAGTCATGGACTAGACGGTCCCTTATGGTAAGAACCAGTCTTT
GTTCAGAGATTCTAACTAAGGACCTTAGAACTGATGGCTTGAGCGAGTGCTCAATAAC
CAAATATTTTTCGCATCTGCACCTTTTATTAAGATGTCAGCTTTACAAATCGCATACGA
TTCATCGTTAATCTCTTGCGCAAAAAATTCAAGGTCAGCTGTAGAATTCAGCTCTTTTA
AATACTCTTGGGCAACGGAACCCATACCACCTGTTCAAGCAGCTTCTAACTAAGGATC
CAAATTTGAATGGGCTGTGGCTAATAACAAGGATGTCATCAATATGAGCTTAGGCGG
ACCAAGCAGTTCAAGAGCTCTTAAAAATGCCGTGGATACAGCAAATAACCGTGGAGT
CGTTGTTTGTGCGGCCCGCGGGT

Appendix II

Blast results of protease gene

PG05 (1101bp gene)

gb|AY817143.1| Bacillus licheniformis strain RG2 keratinolytic protease gene,
Score = 1088 bits (1206), Expect = 0.0 Identities = 689/732 (**94**%), Gaps =
13/732 (1%)Strand=Plus/Plus

```
PG05  375   TCATGGATACAGGAATCCAAGCTTCTCATCCGGACTTGAACGTATTCGGCGGACCAAGCT  434
KLPG  404   TCCTGGATACAGGAATCCAAGCTTCTCATCCGGACTTGAACGTAGTCGGCGGAGCAAGCT  463

PG05  435   TTGTGGCTGGGGAAGCTTATAACACCGACGGCAACGGACACGGCACACATGTTGCCGGTA  494
KLPG  464   TTGTGGCTGGCGAAGCTTATAACACCGACGGCAACGGACACGGCACACATGTTGCCGGTA  523

PG05  495   CAGTAGCTGCGCTTGACAATACAACGGGTGTATTAGGCGTTGCGCCAAGCGTATCCTTGT  554
KLPG  524   CAGTAGCTGCGCTTGACAATACAACGGGTGTATTAGGCGTTGCGCCAAGCGTATCCTTGT  583

PG05  555   ACGCGGTTAAAGTACTGAATTCAAGCGGAAGCGGATCATACAGCGGCATTGTAAGCGGTA  614
KLPG  584   ACGCGGTTAAAGTACTGAATTCAAGCGGAAGCGGATCATACAGCGGCATTGTAAGCGGAA  643

PG05  615   TCCAGTGGGCGACAACAAACGGCATGGATGTTATCAATATGAGCCTTGGGGGAGCATCAC  674
KLPG  644   TCGAGTGGGCGACAACAAACGGCATGGATGTTATCAATATGAGCCTTGGGGGAGCATCAG  703

PG05  675   GCTCAACAGCGATGAAACAGGCAGTCGACAATGCATATGCAAGAGGGGTTGTCGTTGTAG  734
KLPG  704   GCTCGACAGCGATGAAACAGGCAGTCGACAATGCATATGCAAGAGGGGTTGTCGTTGTAG  763

PG05  735   CTGCAGCAGGGAACAGCGGATCTTCCGGAAAACACGAATACAATTGGCTATCCTGCAAAA  794
KLPG  764   CTGCAGCAGGGAACAGCGGATCTTCAGG-AAACACGAATACAATTGGCTATCCTGCGAAA  822

PG05  795   TACGATTCTGTCATCGCAATTGGTGCGGTATACTCTAACAGCAACAGAGCTTCATTTTCC  854
KLPG  823   TACGATTCTGTCATCGCTGTTGGTGCGGTAGACTCTAACAGCAACAGAGCTTCATTTTCC  882

PG05  855   AGTGTGGGAGCAGAGCTTGAAGTCATGGCTCCTGGCGCAGGTGTATACAGCTCTTACCCC  914
KLPG  883   AGTGTGGGAGCAGAGCTTGAAGTCATGGCTCCTGGCGCAGGCGTATACAGCACTTA-CCC  941

PG05  915   ACCCAACACTTATGCAACATTGCACGGAACGTCAATGGCATTCTCCTCATGTAGCCGGGA  974
KLPG  942   AACGAACACTTATGCAACATTGAACGGAACGTCAATGGC-TTCTCCTCATGTAG-CGGGA  999

PG05  975   GCA-CAGCTATGATCTTGTCAAAACATCCCAATCTTTCAGCTTCACGAGTCCGACAAGCG  1033
KLPG  1000  GCAGCAGCTTTGATCTTGTCAAAACATCCGAACCTTTCAGCTTCACAAGTCCG-CAACCG  1058

PG05  1034  TCTCTC--ACACGGCTACTTATTT-GCAAGCTCCTGCTACATAATCGAAAAGGTACTGAT  1090
KLPG  1059  TCTCTCCAGCACGGCGACTTATTTGGGAAGCTCCTTCTAC--TATGGGAAAGGT-CTGAT  1115

PG05  1091  CAATGTC-AAGC  1101
KLPG  1116  CAATGTCGAAGC  1127
```

*KLPG: keratinolytic protease gene

PG08(1156 bp gene)

gb|EU090908.1| Bacillus licheniformis strain YP1A protease gene,
Score = 1160 bits (1286), Expect = 0.0,Identities = 770/825 (**93%**), Gaps = 27/825 (3%),Strand=Plus/Plus

```
PG08  303   ACGGGCTTCCTCGCATTAAAGCGGAGAAAAAGCAGGCTCTCGGCTTTAAAGGAACGAATG  362
BLPG  332   ACGGCGTTCCTCTCATTAAAGCGGACAAAGTGCAGGCTCAAGGCTTTAAGGGAGCGAATG  391

PG08  363   TAAAAGTAGCCGTCCTGGATACAGGAATCCAAGCTTCTCATCCGGACTTGAACGTAGTCG  422
BLPG  392   TAAAAGTAGCCGTCCTGGATACAGGAATCCAAGCTTCTCATCCGGACTTGAACGTAGTCG  451

PG08  423   GCGGAGCAAGCTTTGTGGCTGGCGAAGCTTATAACACCGACGGCAACGGACACGGCACAC  482
BLPG  452   GCGGAGCAAGCTTTGTGGCTGGCGAAGCTTATAACACCGACGGCAACGGACACGGCACAC  511

PG08  483   ATGTTGCCGGTACAGTAGCTGCGCTTGACAATACAACGGGTGTATTAGGCGTTGCGCCAA  542
BLPG  512   ATGTTGCCGGTACAGTAGCTGCGCTTGACAATACAACGGGTGTATTAGGCGTTGCGCCAA  571

PG08  543   GCGTATCCTTGTACGCGGTTAAAGTACTGAATTCAAGCGGAAGCGGATCATACAGCGGCA  602
BLPG  572   GCGTATCCTTGTACGCGGTTAAAGTACTGAATTCAAGCGGAAGCGGATCATACAGCGGCA  631

PG08  603   TTGTAAGCGGTATCGAGTGGGCGACAACAAACGGCATGGATGTTATCAATATGAGCCTTG  662
BLPG  632   TTGTAAGCGGAATCGAGTGGGCGACAACAAACGGCATGGATGTTATCAATATGAGCCTTG  691

PG08  663   GGGGAGCATCAGGCTCGACAGCGATGAAACAGGCAGTCGACAATGCATATGCAAGAGGGG  722
BLPG  692   GGGGAGCATCAGGCTCGACAGCGATGAAACAGGCAGTCGACAATGCATATGCAAGAGGGG  751

PG08  723   TTGTCGTTGTAGCTGCAGCAGGGAACAGCGGATCTTCAGGAAAACACGAATACAATTGGC  782
BLPG  752   TTGTCGTTGTAGCTGCAGCAGGGAACAGCGGATCTTCAGG-AAACACGAATACAATTGGC  810

PG08  783   TATCCTGCGAAATACGATTCTGTCATCGCAGTTGGGTGCGGTAGACTCTAACAGCAACAG  842
BLPG  811   TATCCTGCGAAATACGATTCTGTCATCGCTGTT-GGCGCGGTAGACTCTAACAGCAACAG  869

PG08  843   AGCTTCATTTTCCAGTGGTGGGGAGCAGAGCTTGAAGTCATGGCTCCTGGGCGCAAGGCG  902
BLPG  870   AGCTTCATTTTCCAGTG--TGGGAGCAGAGCTTGAAGTCATGGCTCCT-GGCGC-AGGCG  925

PG08  903   TATACAGCACTTACCAACCGAACACTTATGCAACATTGGAACGGAACCGTCAATGGCTTC  962
BLPG  926   TATACAGCACTTACCCAACGAACACTTATGCAACATT-GAACGGAA-CGTCAATGGCTTC  983

PG08  963   CTCTCATGTAGCGGAAGCAGCAGCTTTTGATCTTTGTCAAAAACATCCGAATCTTTCAGC  1022
BLPG  984   TCCTCATGTAGCGGGAGCAGCAGC-TTTGATC-TTGTC-AAAACATCCGAACCTTTCAGC  1040

PG08  1023  TTCACCAAGTTCCGGCAAACGGTCTTCTTTCAGCAACGAGCGGACCTTATTTGGGAAGGC  1082
BLPG  1041  TTCA-CAAG-TCCG--CAACCGTCT--CTCCAGC-ACG-GCG--ACTTATTTGGGAA-GC  1089

PG08  1083  TCCGTTCTACTAAGGAAAAGGTCCTGAATCGATTGTCGAAAGCTG  1127
BLPG  1090  TCC-TTCTACTATGGGAAAGGT-CTG-ATC-AATGTCG-AAGCTG  1129
```

*BLPG: Bacillus licheniformis protease gene

PG20 (819 bp gene)

gb|AY743586.1| Bacillus pumilus strain 115b organic solvent tolerant protease gene, Score = 1359 bits (1506), Expect = 0.0, Identities = 761/766 (**99**%), Gaps = 0/766 (0%), Strand=Plus/Plus

```
PG20  38    ATCAAGGTGCTAATGTCAAAATAGCCGTCCTTGATACTGGAATCCACGCTGCACACCCTG  97
OSTP  386   ATAAAGGTGCTAATGTCAAAGTAGCCGTCCTTGATACTGGAATCCACGCTGCACACCCTG  445

PG20  98    ACTTAAATGTTGCAGGCGGTGCGAGCTTCGTCCCTTCAGAGCCAAATGCCACCCAAGACT  157
OSTP  446   ACTTAAATGTTGCAGGCGGTGCGAGCTTCGTCCCTTCAGAGCCAAATGCCACCCAAGACT  505

PG20  158   TTCAATCACATGGAACTCACGTAGCTGGAACCATTGCTGCCCTTGATAACACAATTGGTG  217
OSTP  506   TTCAATCACATGGAACTCACGTAGCTGGAACCATTGCTGCCCTTGATAACACAATTGGTG  565

PG20  218   TTCTTGGGGTTGCTCCAAACGCTTCCCTATATGCTGTGAAAGTATTAGACCGTAATGGTG  277
OSTP  566   TTCTTGGGGTTGCTCCAAACGCTTCCCTATATGCTGTGAAAGTATTAGACCGTAATGGTG  625

PG20  278   ACGGACAATACAGCTGGATTATTAGCGGTATTGAATGGGCTGTAGCTAATAATATGGATG  337
OSTP  626   ACGGACAATACAGCTGGATTATTAGCGGTATTGAATGGGCTGTAGCTAATAATATGGATG  685

PG20  338   TCATCAATATGAGCTTAGGCGGACCAAGCGGTTCAACAGCGCTTAAAAATGCCGTGGATA  397
OSTP  686   TCATCAATATGAGCTTAGGCGGACCAAGCGGTTCAACAGCGCTTAAAAATGCCGTGGATA  745

PG20  398   CAGCGAATAACCGTGGAGTCGTTGTTGTGGCGGCCGCAGGTAATTCTGGCTCTAGTGGCT  457
OSTP  746   CAGCGAATAATCGTGGAGTCGTTGTTGTGGCGGCCGCAGGTAATTCTGGCTCTAGTGGCT  805

PG20  458   CTAGAAGTACAGTTGGCTATCCAGCAAAATACGATTCTACAATTGCCGTTGCCAATGTAA  517
OSTP  806   CTAGAAGTACAGTTGGCTATCCTGCAAAATACGATTCTACAATTGCCGTTGCCAATGTAA  865

PG20  518   ACAGTAACAATGTCAGAAACTCATCTTCTAGCGCAGGTCCTGAATTAGATGTTTCTGCAC  577
OSTP  866   ACAGTAGCAATGTCAGAAACTCATCTTCTAGCGCAGGTCCTGAATTAGATGTTTCTGCAC  925

PG20  578   CTGGTACTTCTATTTTAAGTACAGTCCCAAGCAGCGGATACACATCTTATACTGGAACAT  637
OSTP  926   CTGGTACTTCTATTTTAAGTACAGTCCCAAGCAGCGGATACACATCTTATACTGGAACAT  985

PG20  638   CTATGGCGTCTCCTCATGTAGCAGGAGCAGCAGCGCTTATTCTTTCTAAAAACCCGAATC  697
OSTP  986   CTATGGCGTCTCCTCATGTAGCAGGAGCAGCAGCGCTTATTCTTTCTAAAAACCCGAATC  1045

PG20  698   TAACAAATTCACAGGTTCGCCAGCGCTTAGAAAATACAGCAACACCGCTTGGTGACTCAT  757
OSTP  1046  TAACAAATTCACAGGTTCGCCAGCGCTTAGAAAATACAGCAACACCGCTTGGTGACTCAT  1105

PG20  758   TCTATTACGGAAAAGGGTTAATCAACGTTCAAGCAGCTTCTAACTA  803
OSTP  1106  TCTATTACGGAAAAGGGTTAATCAACGTTCAAGCAGCTTCTAACTA  1151
```

*OSTP: organic solvent tolerant protease

PG23(950 bp gene)

gb|FJ584420.1| Bacillus pumilus strain TMS55 alkaline serine protease precursor gene, Score = 244 bits (270), Expect = 5e-61, Identities = 502/738 (**68**%), Gaps = 9/738 (1%), Strand=Plus/Plus

```
PG23  42    TAAAGGCGTTAATGGCAAAGTACTGGTTGTTGATTTTGCCTATAATGCTGCACATCCTGG  101
AKSP  549   TAAAGGTGCTAATGTCAAAGTAGCTGTCCTTGATACTGGAATCCACGCTGCACACCCTGA  608

PG23  102   CTTGAATGATGCATGGGCTGCGAGGTTCCCTGCGTGAAATTGAAATGCCCTCCCAAATTT  161
AKSP  609   CTTAAATGCTGCAGGCGGTGCGAGCTTCGTCCCTTCAGAGCCAAATGCCACCCAAGACTT  668

PG23  162   TCACTCCCTTGGAACTCGAATAATTGGAAGCATTGCTGGCCTTGCCTACTGAGTGGGCGT  221
AKSP  669   TCAATCACATGGAACTCACGTAGCTGGAACCATTGCTGCCCTTGATAACACAATTGGTGT  728

PG23  222   TCTTGGGGAATTGCGTTG-GCTTCCCTGCCCGCTGTCCAAGCATTAAACCTGTCTGGCCA  280
AKSP  729   TCTTGGGGTCGCTCCAAGCGCTTCCCTATATGCTGTGAAAGCATTAGACCGCAATGGCGA  788

PG23  281   CCTGCATTAATGATGGATTATTAGCGCGATTGAAGGGGGTGTGCGCAATGAGAGGTATGT  340
AKSP  789   CGGACAATACAGCTGGATTATTAGCGGTATTGAATGGGCTGTAGCCAATAACATGGATGT  848

PG23  341   CGTTTCTATGATTTTTGACTCAGTGCGCTCGTCCGCACGGCTGAAAAAAGCGGTTTATAC  400
AKSP  849   CATCAATATGAGTTTAGGCGGAGCAAGCGGTTCAACAGCACTTAAAAATGCCGTTGATAC  908

PG23  401   TCACTAAAAGG-GCGGAATACTTGTATCTGCCGAATCAGGGAATTACGGCTCAAAGAACT  459
AKSP  909   AG-CAAACAGCCGCGGAGTCGTTGCTGTTGCCGCCGCAGGTAATTCAGGCTCTAGTGGCT  967

PG23  460   CTAGAACAAAAGGTGGCTATCCAGCCAAAAACCAGTCTAAGGTTGCTGTGCTGGATGTTT  519
AKSP  968   CTAGAAGTACAGTTGGCTATCCAGCCAAATACGAGTCTACAATTGCTGTTGCCAATGTAA  1027

PG23  520   TCCGTAACCTCGTCCCCCCTGAATCTTCTACCGAAATCCCTGATTTAGATGTTTCTGCCC  579
AKSP  1028  ACAGTAACAATGTCAGAAACTCATCTTCTAGCGCAGGTCCTGAATTAGATGTTTCTGCAC  1087

PG23  580   CTGCCACTTCTATTTTTAATAATACGCCGAGTTTCGCACACACTTCTCATACGGGAACTC  639
AKSP  1088  CTGGTACTTCTATTTTAAGTACAGTGCCAAGCAGTGGATACACTTCTTATACTGGAACAT  1147

PG23  640   CTCTG--GTCTCACCCTGTCGCTTACGAGAAACCTGTCTTACTTTTTCTAATAACCAAAA  697
AKSP  1148  CTATGGCGTCTCCTCATGTAGC--AGGAGCAGCAGCGCTTATTCTTTCTAAAAACCCGAA  1205

PG23  698   CCTAACCTATCTCCAAGCTCACGAGCGCTGTATAAAATACTCCGACACCGCTTGGTGACT  757
AKSP  1206  CCTAACAAATTCACAGGTTCGCCAGCGCT-TAGAAAATACAGCGACACCGCTTGGTGACT  1264

PG23  758   CATTCTATTATGTGAAAA  775
AKSP  1265  CATTCTATTATG-GAAAA  1281
```

*AKSP: alkaline serine protease

PG25 (810bp gene)

gb|AY743586.1| Bacillus pumilus strain 115b organic solvent tolerant protease gene, Score = 1357 bits (1504), Expect = 0.0, Identities = 764/772 (**98**%), Gaps = 0/772 (0%), Strand=Plus/Plus

```
PG25  21    GCTAAAAGTTTTAAGGGTGCAAATGTCAAAGTAGCCGTCCTTGATACTGGAATCCACGCT  80
OSTP  376   GCTCAAGGTTATAAAGGTGCTAATGTCAAAGTAGCCGTCCTTGATACTGGAATCCACGCT  435

PG25  81    GCACACCCTGACTTAAATGTTGCAGGCGGTGCGAGCTTCGTCCCTTCAGAGCCAAATGCC  140
OSTP  436   GCACACCCTGACTTAAATGTTGCAGGCGGTGCGAGCTTCGTCCCTTCAGAGCCAAATGCC  495

PG25  141   ACCCAAGACTTTCAATCACATGGAACTCACGTAGCTGGAACCATTGCTGCCCTTGATAAC  200
OSTP  496   ACCCAAGACTTTCAATCACATGGAACTCACGTAGCTGGAACCATTGCTGCCCTTGATAAC  555

PG25  201   ACAATTGGTGTTCTTGGGGTTGCTCCAAACGCTTCCCTATATGCTGTGAAAGTATTAGAC  260
OSTP  556   ACAATTGGTGTTCTTGGGGTTGCTCCAAACGCTTCCCTATATGCTGTGAAAGTATTAGAC  615

PG25  261   CGTAATGGTGACGGACAATACAGCTGGATTATTAGCGGTATTGAATGGGCTGTAGCTAAT  320
OSTP  616   CGTAATGGTGACGGACAATACAGCTGGATTATTAGCGGTATTGAATGGGCTGTAGCTAAT  675

PG25  321   AATATGGATGTCATCAATATGAGCTTAGGCGGACCAAGCGGTTCAACAGCGCTTAAAAAT  380
OSTP  676   AATATGGATGTCATCAATATGAGCTTAGGCGGACCAAGCGGTTCAACAGCGCTTAAAAAT  735

PG25  381   GCCGTGGATACAGCGAATAACCGTGGAGTCGTTGTTGTGGCGGCCGCAGGTAATTCTGGC  440
OSTP  736   GCCGTGGATACAGCGAATAATCGTGGAGTCGTTGTTGTGGCGGCCGCAGGTAATTCTGGC  795

PG25  441   TCTAGTGGCTCTAGAAGTACAGTTGGCTATCCAGCAAAATACGATTCTACAATTGCCGTT  500
OSTP  796   TCTAGTGGCTCTAGAAGTACAGTTGGCTATCCTGCAAAATACGATTCTACAATTGCCGTT  855

PG25  501   GCCAATGTAAACAGTAACAATGTCAGAAACTCATCTTCTAGCGCAGGTCCTGAATTAGAT  560
OSTP  856   GCCAATGTAAACAGTAGCAATGTCAGAAACTCATCTTCTAGCGCAGGTCCTGAATTAGAT  915

PG25  561   GTTTCTGCACCTGGTACTTCTATTTTAAGTACAGTCCCAAGCAGCGGATACACATCTTAT  620
OSTP  916   GTTTCTGCACCTGGTACTTCTATTTTAAGTACAGTCCCAAGCAGCGGATACACATCTTAT  975

PG25  621   ACTGGAACATCTATGGCGTCTCCTCATGTAGCAGGAGCAGCAGCGCTTATTCTTTCTAAA  680
OSTP  976   ACTGGAACATCTATGGCGTCTCCTCATGTAGCAGGAGCAGCAGCGCTTATTCTTTCTAAA  1035

PG25  681   AACCCGAATCTAACAAATTCACAGGTTCGCCAGCGCTTAGAAAATACAGCAACACCGCTT  740
OSTP  1036  AACCCGAATCTAACAAATTCACAGGTTCGCCAGCGCTTAGAAAATACAGCAACACCGCTT  1095

PG25  741   GGTGACTCATTCTATTACGGAAAAGGGTTAATCAACGTTCAAGCAGCTTCTA  792
OSTP  1096  GGTGACTCATTCTATTACGGAAAAGGGTTAATCAACGTTCAAGCAGCTTCTA  1147
```

*OSTP: organic solvent tolerant protease

PG-SC (429 bp gene)

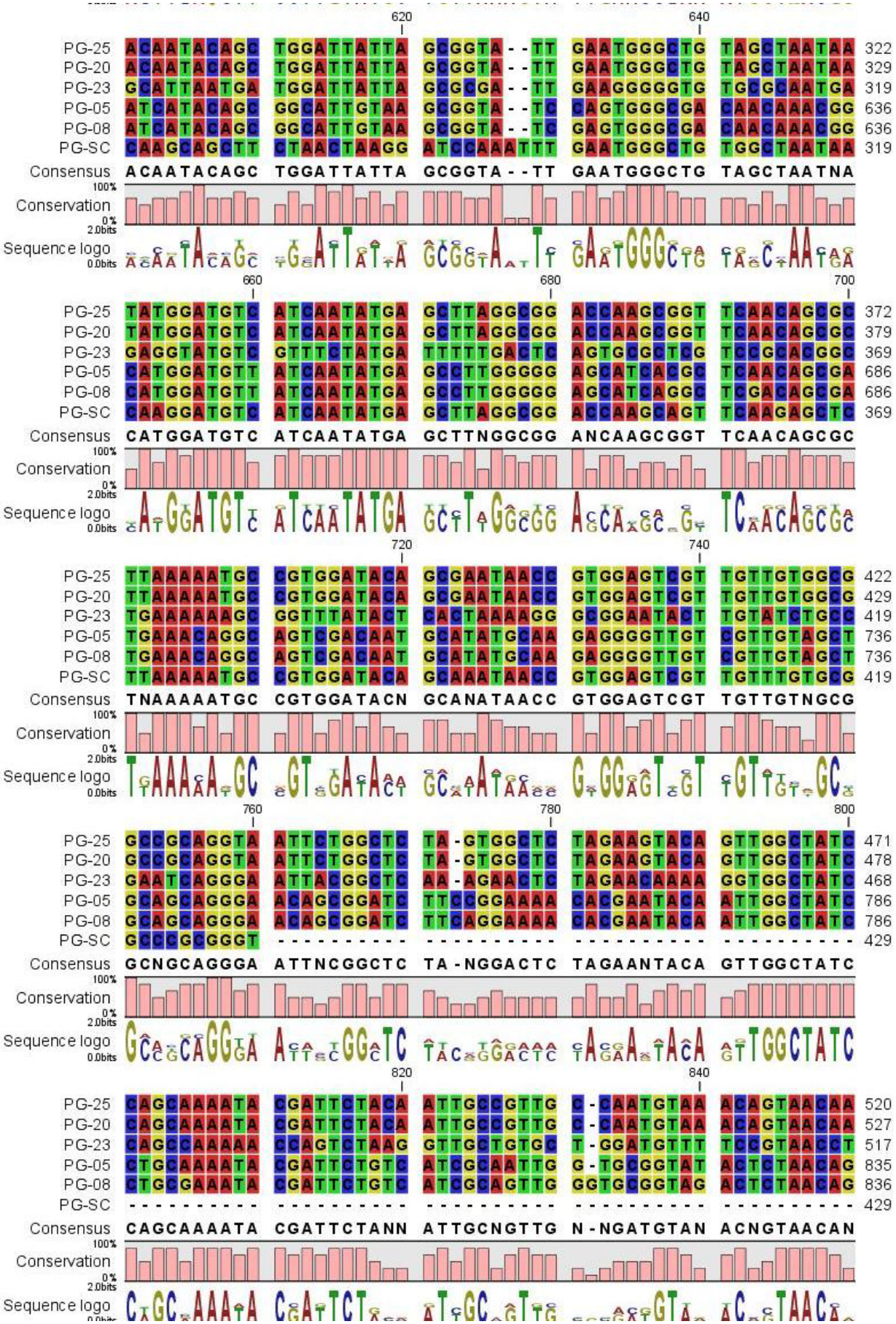
gb|AY743586.1| Bacillus pumilus strain 115b organic solvent tolerant protease gene, Score = 176 bits (194), Expect = 1e-40, Identities = 114/125 (**91**%), Gaps = 0/125 (0%),Strand=Plus/Plus

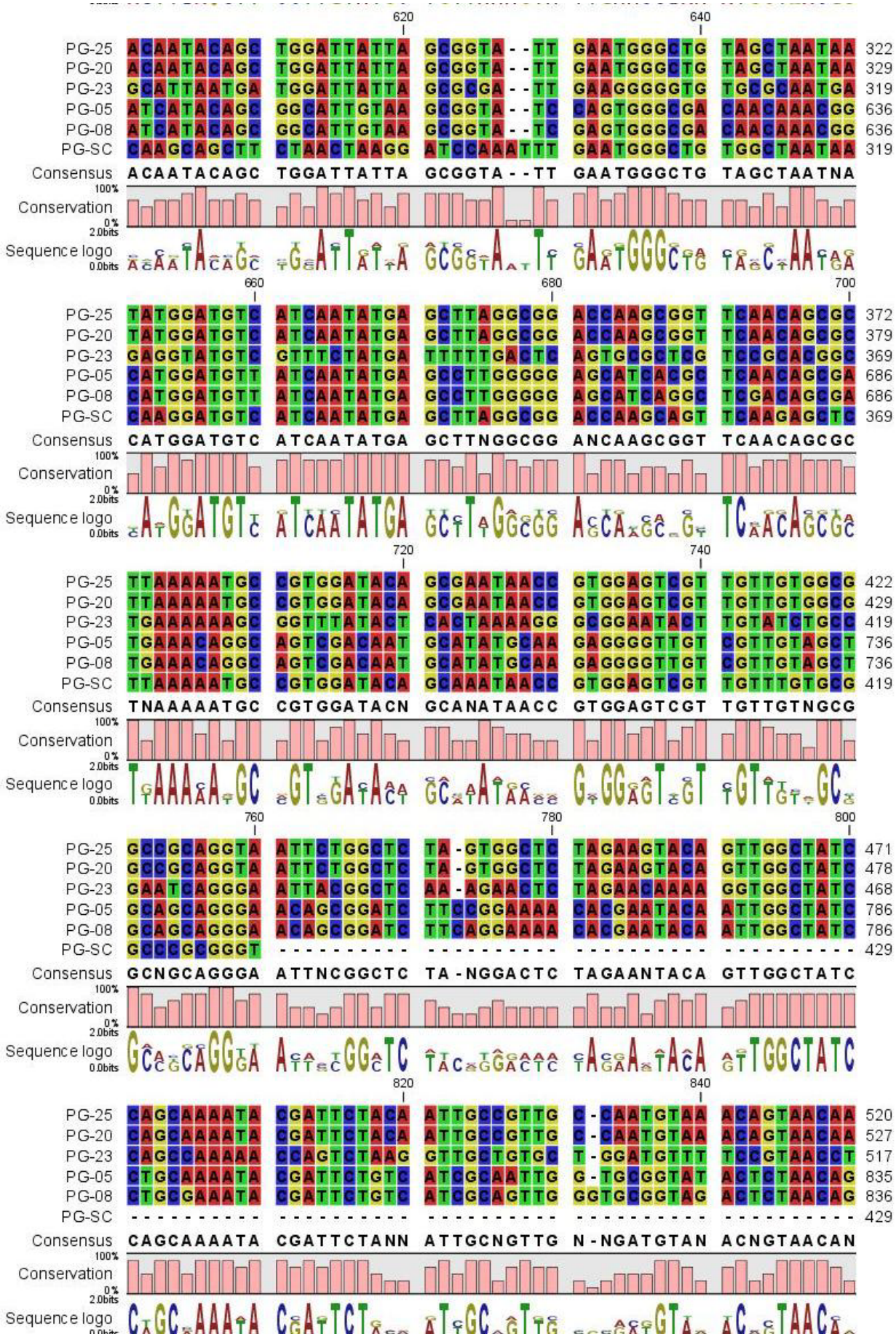
```
PGSc  298  TTGAATGGGCTGTGGCTAATAACAAGGATGTCATCAATATGAGCTTAGGCGGACCAAGCA  357
OSTP  656  TTGAATGGGCTGTAGCTAATAATATGGATGTCATCAATATGAGCTTAGGCGGACCAAGCG  715

PGSc  358  GTTCAAGAGCTCTTAAAAATGCCGTGGATACAGCAAATAACCGTGGAGTCGTTGTTTGTG  417
OSTP  716  GTTCAACAGCGCTTAAAAATGCCGTGGATACAGCGAATAATCGTGGAGTCGTTGTTGTGG  775

PGSc  418  CGGCC  422
OSTP  776  CGGCC  780
```

*OSTP: organic solvent tolerant protease

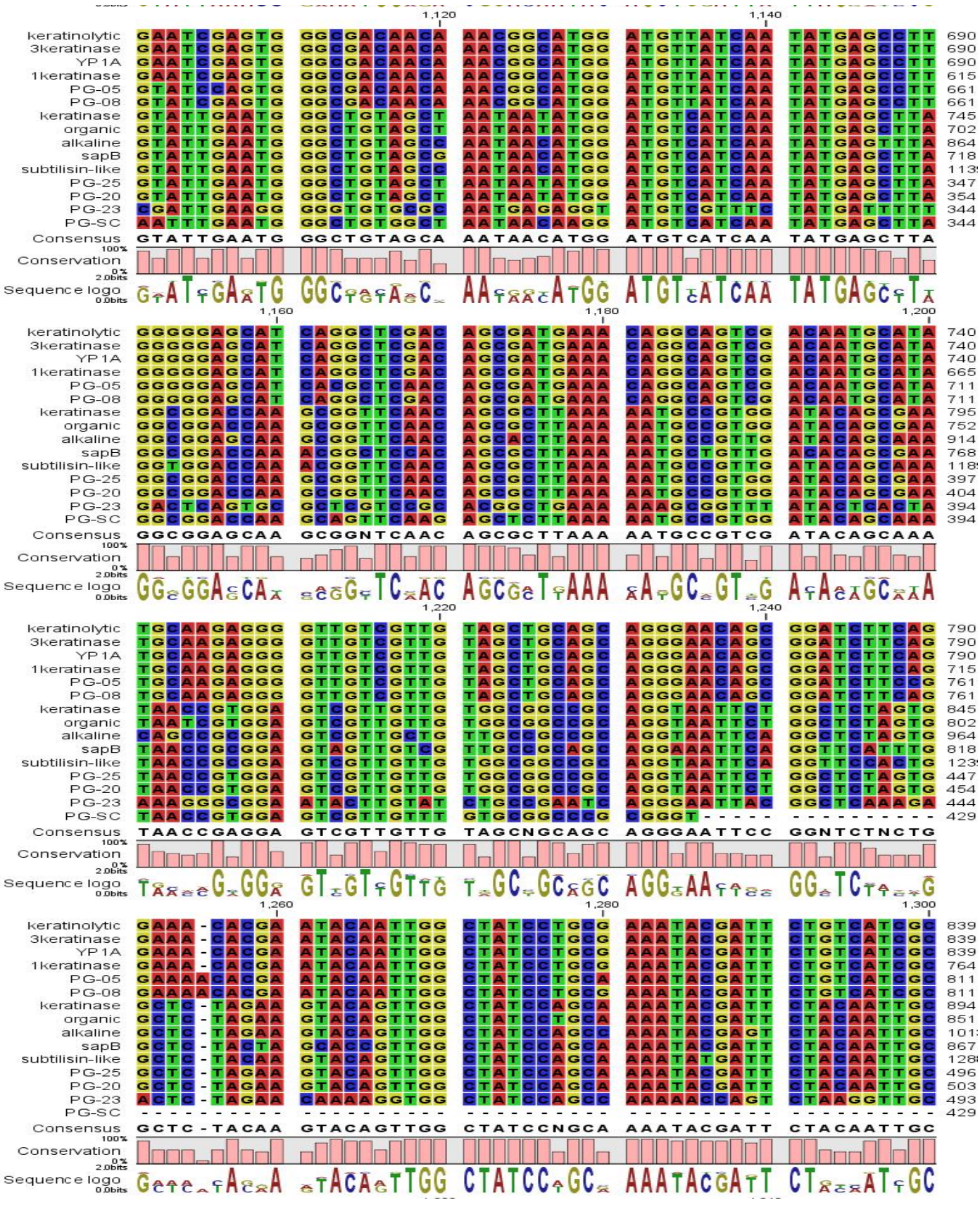
Appendix III

Multiple alignment between sequenced protease gene from the bacteria under study as well as from soil sample.




620
640
PG-25 ACAATACAGC TGGATTATTA GCGGTA--TT GAATGGGCTG TAGCTAATAA 322
PG-20 ACAATACAGC TGGATTATTA GCGGTA--TT GAATGGGCTG TAGCTAATAA 329
PG-23 GCATTAATGA TGGATTATTA GCGCGA--TT GAAGGGGGTG TGCGCAATGA 319
PG-05 ATCATACAGC GGCATTGTAA GCGGTA--TC CAGTGGGCGA CAACAAACGG 636
PG-08 ATCATACAGC GGCATTGTAA GCGGTA--TC GAGTGGGCGA CAACAAACGG 636
PG-SC CAAGCAGCTT CTAACTAAGG ATCCAAATTT GAATGGGCTG TGGCTAATAA 319
Consensus ACAATACAGC TGGATTATTA GCGGTA--TT GAATGGGCTG TAGCTAATNA
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits
660
680
700
PG-25 TATGGATGTC ATCAATATGA GCTTAGGCGG ACCAAGCGGT TCAACAGCGC 372
PG-20 TATGGATGTC ATCAATATGA GCTTAGGCGG ACCAAGCGGT TCAACAGCGC 379
PG-23 GAGGTATGTC GTTTCTATGA TTTTTGACTC AGTGCGCTCG TCCGCACGGC 369
PG-05 CATGGATGTT ATCAATATGA GCCTTGGGGG AGCATCACGC TCAACAGCGA 686
PG-08 CATGGATGTT ATCAATATGA GCCTTGGGGG AGCATCAGGC TCGACAGCGA 686
PG-SC CAAGGATGTC ATCAATATGA GCTTAGGCGG ACCAAGCAGT TCAAGAGCTC 369
Consensus CATGGATGTC ATCAATATGA GCTTNGGCGG ANCAAGCGGT TCAACAGCGC
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits
720
740
PG-25 TTAAAAAATGC CGTGGATACA GCGAATAACC GTGGAGTCGT TGTTGTGGCG 422
PG-20 TTAAAAAATGC CGTGGATACA GCGAATAACC GTGGAGTCGT TGTTGTGGCG 429
PG-23 TGAAAAAAGC GGTTTATACT CACTAAAAGG GCGGAATACT TGTATCTGCC 419
PG-05 TGAAACAGGC AGTCGACAAT GCATATGCAA GAGGGGTTGT CGTTGTAGCT 736
PG-08 TGAAACAGGC AGTCGACAAT GCATATGCAA GAGGGGTTGT CGTTGTAGCT 736
PG-SC TTAAAAATGC CGTGGATACA GCAAATAACC GTGGAGTCGT TGTTTGTGCG 419
Consensus TNAAAAATGC CGTGGATACN GCANATAACC GTGGAGTCGT TGTTGTNGCG
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits
760
780
800
PG-25 GCCGCAGGTA ATTCTGGCTC TA-GTGGCTC TAGAAGTACA GTTGGCTATC 471
PG-20 GCCGCAGGTA ATTCTGGCTC TA-GTGGCTC TAGAAGTACA GTTGGCTATC 478
PG-23 GAATCAGGGA ATTACGGCTC AA-AGAACTC TAGAACAAAA GGTGGCTATC 468
PG-05 GCAGCAGGGA ACAGCGGATC TTCCGGAAAA CACGAATACA ATTGGCTATC 786
PG-08 GCAGCAGGGA ACAGCGGATC TTCAGGAAAA CACGAATACA ATTGGCTATC 786
PG-SC GCCCGCGGGT ---------- ---------- ---------- ---------- 429
Consensus GCNGCAGGGA ATTNCGGCTC TA-NGGACTC TAGAANTACA GTTGGCTATC
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits
820
840
PG-25 CAGCAAAATA CGATTCTACA ATTGCCGTTG C-CAATGTAA ACAGTAACAA 520
PG-20 CAGCAAAATA CGATTCTACA ATTGCCGTTG C-CAATGTAA ACAGTAACAA 527
PG-23 CAGCCAAAAA CCAGTCTAAG GTTGCTGTGC T-GGATGTTT TCCGTAACCT 517
PG-05 CTGCAAAATA CGATTCTGTC ATCGCAATTG G-TGCGGTAT ACTCTAACAG 835
PG-08 CTGCGAAATA CGATTGTGTC ATCGCAGTTG GGTGCGGTAG ACTCTAACAG 836
PG-SC ---------- ---------- ---------- ---------- ---------- 429
Consensus CAGCAAAATA CGATTCTANN ATTGCNGTTG N-NGATGTAN ACNGTAACAN
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits

Multiple sequence alignment of protease gene sequences from bacteria and soil sample under study and protease sequences from other homologous species.



Appendix IV
Protein blast results

PG05

Frame 2 ----360 aa---*B. amyloliqueficens*

```
sp|P00780.1|SUBT_BACLI  RecName: Full=Subtilisin Carlsberg; Flags: Precursor
 emb|CAB56500.1|  subC [Bacillus licheniformis]
Length=379

 Score =  253 bits (646),  Expect = 2e-65, Method: Compositional matrix adjust.
 Identities = 152/231 (65%), Positives = 169/231 (73%), Gaps = 17/231 (7%)

PG05  116  QIKSRFMDTGIQASHPDLNVFGGPSFVAGEAYNTDGNGHGTHVAGTVAALDNTTGVLGVA  175
           +K   +DTGIQASHPDLNV GG SFVAGEAYNTDGNGHGTHVAGTVAALDNTTGVLGVA
SubC  130  NVKVAVLDTGIQASHPDLNVVGGASFVAGEAYNTDGNGHGTHVAGTVAALDNTTGVLGVA  189

PG05  176  PSVSLYAVKVLNSSGSGSYSGIVSGIQWATTNGMDVINMSLGGASRSTAMKQAVDNAYAR  235
           PSVSLYAVKVLNSSGSG+YSGIVSGI+WATTNGMDVINMSLGG S STAMKQAVDNAYAR
SubC  190  PSVSLYAVKVLNSSGSGTYSGIVSGIEWATTNGMDVINMSLGGPSGSTAMKQAVDNAYAR  249

PG05  236  GVVVVAAAGNSGSSGKHEYNWLSCKIRFCHRNWCGILQQQSFIFQCGSRASHGSWRRCIQ  295
           GVVVVAAAGNSGSSG    N +    ++      G +   S      +RAS  S    ++
SubC  250  GVVVVAAAGNSGSSGNT--NTIGYPAKYDSVIAVGAVDSNS------NRASFSSVGAELE  301

PG05  296  LLP--------HPT-LMQHCTERQWHSPHVAGSTAMILSKHPNLSASRVRQ  337
           ++         +PT            SPHVAG+ A+ILSKHPNLSAS+VR
SubC  302  VMAPGAGVYSTYPTSTYATLNGTSMASPHVAGAAALILSKHPNLSASQVRN  352
```

*SubC: Subtilisin Carlsberg

PG08

Frame 2--- 279aa— *S. arlettae.*

```
sp|P00780.1|SUBT_BACLI  RecName: Full=Subtilisin Carlsberg; Flags: Precursor
 emb|CAB56500.1|  subC [Bacillus licheniformis]
Length=379

 Score =  280 bits (716),  Expect = 2e-73, Method: Compositional matrix
adjust.
 Identities = 169/256 (66%), Positives = 185/256 (72%), Gaps = 7/256 (2%)

PG08  78   KRLPETTKKDRGQIA---PEIQPNGLPRIKAEKKQALGFKGTNVKVAVLDTGIQASHPDL  134
           K  P+    +   +A    +  P G+P IKA+K QA GFKG NVKVAVLDTGIQASHPDL
SubC  88   KNDPDVAYVEEDHVAHALAQTVPYGIPLIKADKVQAQGFKGANVKVAVLDTGIQASHPDL  147

PG08  135  NVVGGASFVAGEAYNTDGNGHGTHVAGTVAALDNTTGVLGVAPSVSLYAVKVLNSSGSGS  194
           NVVGGASFVAGEAYNTDGNGHGTHVAGTVAALDNTTGVLGVAPSVSLYAVKVLNSSGSG+
SubC  148  NVVGGASFVAGEAYNTDGNGHGTHVAGTVAALDNTTGVLGVAPSVSLYAVKVLNSSGSGT  207

PG08  195  YSGIVSGIEWATTNGMDVINMSLGGASGSTAMKQAVDNAYARGVVVVAAAGNSGSSGKHE  254
           YSGIVSGIEWATTNGMDVINMSLGG SGSTAMKQAVDNAYARGVVVVAAAGNSGSSG
Subc  208  YSGIVSGIEWATTNGMDVINMSLGGPSGSTAMKQAVDNAYARGVVVVAAAGNSGSSGNT-  266

PG08  255  YNWLSCEIRFCHRSWVRTL--TATELHFPVVGSRASHGSWAQGVYSTYQPNTYATLERNR  312
            N +    ++     V  +   +    F  VG+     +   GVYSTY  +TYATL
SubC  267  -NTIGYPAKYDSVIAVGAVDSNSNRASFSSVGAELEVMAPGAGVYSTYPTSTYATLNGTS  325

PG08  313  QWLPLMRKQQLLIFVK  328
              P +     LI  K
SubC  326  MASPHVAGAAALILSK  341
```

*SubC: Subtilisin Carlsberg

PG20

Frame 1 ----273aa- *B. licheniformis*

gb|ACM47735.1| keratinase precursor [*Bacillus pumilus]*

Length=383

```
 Score =  505 bits (1300),  Expect = 2e-141, Method: Compositional matrix
adjust.
 Identities = 253/257 (98%), Positives = 255/257 (99%), Gaps = 0/257 (0%)

PG20  11   QCDQGANVKIAVLDTGIHAAHPDLNVAGGASFVPSEPNATQDFQSHGTHVAGTIAALDNT  70
           Q  +GANVK+AVLDTGIHAAHPDLNVAGGASFVPSEPNATQDFQSHGTHVAGTIAALDNT
KTPR  127  QGYKGANVKVAVLDTGIHAAHPDLNVAGGASFVPSEPNATQDFQSHGTHVAGTIAALDNT  186

PG20  71   IGVLGVAPNASLYAVKVLDRNGDGQYSWIISGIEWAVANNMDVINMSLGGPSGSTALKNA  130
           IGVLGVAPNASLYAVKVLDRNGDGQYSWIISGIEWAVANNMDVINMSLGGPSGSTALKNA
KTPR  187  IGVLGVAPNASLYAVKVLDRNGDGQYSWIISGIEWAVANNMDVINMSLGGPSGSTALKNA  246

PG20  131  VDTANNRGVVVVAAAGNSGSSGSRSTVGYPAKYDSTIAVANVNSNNVRNSSSSAGPELDV  190
           VDTANNRGVVVVAAAGNSGSSGSRSTVGYPAKYDSTIAVANVNSNNVRNSSSSAGPELDV
KTPR  247  VDTANNRGVVVVAAAGNSGSSGSRSTVGYPAKYDSTIAVANVNSNNVRNSSSSAGPELDV  306

PG20  191  SAPGTSILSTVPSSGYTSYTGTSMASPHVAGAAALILSKNPNLTNSQVRQRLENTATPLG  250
           SAPGTSILSTVPSSGYTSYTGTSMASPHVAGAAALILSKNPNLTNSQVRQRLENTATPLG
KTPR  307  SAPGTSILSTVPSSGYTSYTGTSMASPHVAGAAALILSKNPNLTNSQVRQRLENTATPLG  366

PG20  251  DSFYYGKGLINVQAASN  267
           DSFYYGKGLINVQAASN
KTPR  367  DSFYYGKGLINVQAASN  383
```

* KTPR: keratinase precursor

PG23

Frame 3—304aa ---*B.licheniformis*

emb|CAO03040.1| serine alkaline protease, preproprotein [Bacillus pumilus]
Length=383

```
 Score =  101 bits (252),  Expect = 7e-20, Method: Compositional matrix
adjust.
 Identities = 70/162 (43%), Positives = 87/162 (53%), Gaps = 3/162 (1%)

PG23  75   ASLPAVQAL--NLSGHLHWIISAIEGGVRNERYVVSMIFDSVRSSARLKKAVYTHKG-GI  131
           ASL AV+ L  N  G   WIIS IE  V N   V++M       S  LK AV T    G+
SAlk  196  ASLYAVKVLDRNGDGQYSWIISGIEWAVANNMDVINMSLGGPNGSTALKNAVDTANNRGV  255

PG23  132  LVSAESGNYGSKNSRTKGGYPAKNQSKVAVLDVFRNLVPPESSTEIPDLDVSAPATSIFN  191
           +V A +GN GS  S +  GYPAK  S +AV +V  N V   SS+  P+LDVSAP TSI +
SAlk  256  VVVAAAGNSGSFGSTSTVGYPAKYDSTIAVANVNGNNVRNSSSSAGPELDVSAPGTSILS  315

PG23  192  NTPSFAHTSHTGTPLVSPCRLRETCLTFSNNQNLTYLQAHER  233
             PS  +TS+TGT + SP       L  S   NL+  Q  +R
SAlk  316  TVPSSGYTSYTGTSMASPHVAGAAALILSKYPNLSTSQVRQR  357
```

*Salk:Serine alkaline protease

PG 25

Frame3---268aa—*B.pumilus*

```
gb|ACM47735.1|  keratinase precursor [Bacillus pumilus]
Length=383

 Score =  507 bits (1305),  Expect = 5e-142, Method: Compositional matrix
adjust.
 Identities = 254/257 (98%), Positives = 256/257 (99%), Gaps = 0/257 (0%)

PG25  6    AKSFKGANVKVAVLDTGIHAAHPDLNVAGGASFVPSEPNATQDFQSHGTHVAGTIAALDN  65
           A+ +KGANVKVAVLDTGIHAAHPDLNVAGGASFVPSEPNATQDFQSHGTHVAGTIAALDN
KTPR  126  AQGYKGANVKVAVLDTGIHAAHPDLNVAGGASFVPSEPNATQDFQSHGTHVAGTIAALDN  185

PG25  66   TIGVLGVAPNASLYAVKVLDRNGDGQYSWIISGIEWAVANNMDVINMSLGGPSGSTALKN  125
           TIGVLGVAPNASLYAVKVLDRNGDGQYSWIISGIEWAVANNMDVINMSLGGPSGSTALKN
KTPR  186  TIGVLGVAPNASLYAVKVLDRNGDGQYSWIISGIEWAVANNMDVINMSLGGPSGSTALKN  245

PG25  126  AVDTANNRGVVVVAAAGNSGSSGSRSTVGYPAKYDSTIAVANVNSNNVRNSSSSAGPELD  185
           AVDTANNRGVVVVAAAGNSGSSGSRSTVGYPAKYDSTIAVANVNSNNVRNSSSSAGPELD
KTPR  246  AVDTANNRGVVVVAAAGNSGSSGSRSTVGYPAKYDSTIAVANVNSNNVRNSSSSAGPELD  305

PG25  186  VSAPGTSILSTVPSSGYTSYTGTSMASPHVAGAAALILSKNPNLTNSQVRQRLENTATPL  245
           VSAPGTSILSTVPSSGYTSYTGTSMASPHVAGAAALILSKNPNLTNSQVRQRLENTATPL
KTPR  306  VSAPGTSILSTVPSSGYTSYTGTSMASPHVAGAAALILSKNPNLTNSQVRQRLENTATPL  365

PG25  246  GDSFYYGKGLINVQAAS  262
           GDSFYYGKGLINVQAAS
KTPR  366  GDSFYYGKGLINVQAAS  382
```

*KTPR: keratinase precursor

PGSc

frame3---137aa--- Protease from Soil Metagenomic DNA

ref|YP_001486216.1| subtilisin [Bacillus pumilus SAFR-032]
gb|ABV61656.1| subtilisin [Bacillus pumilus SAFR-032]
Length=381

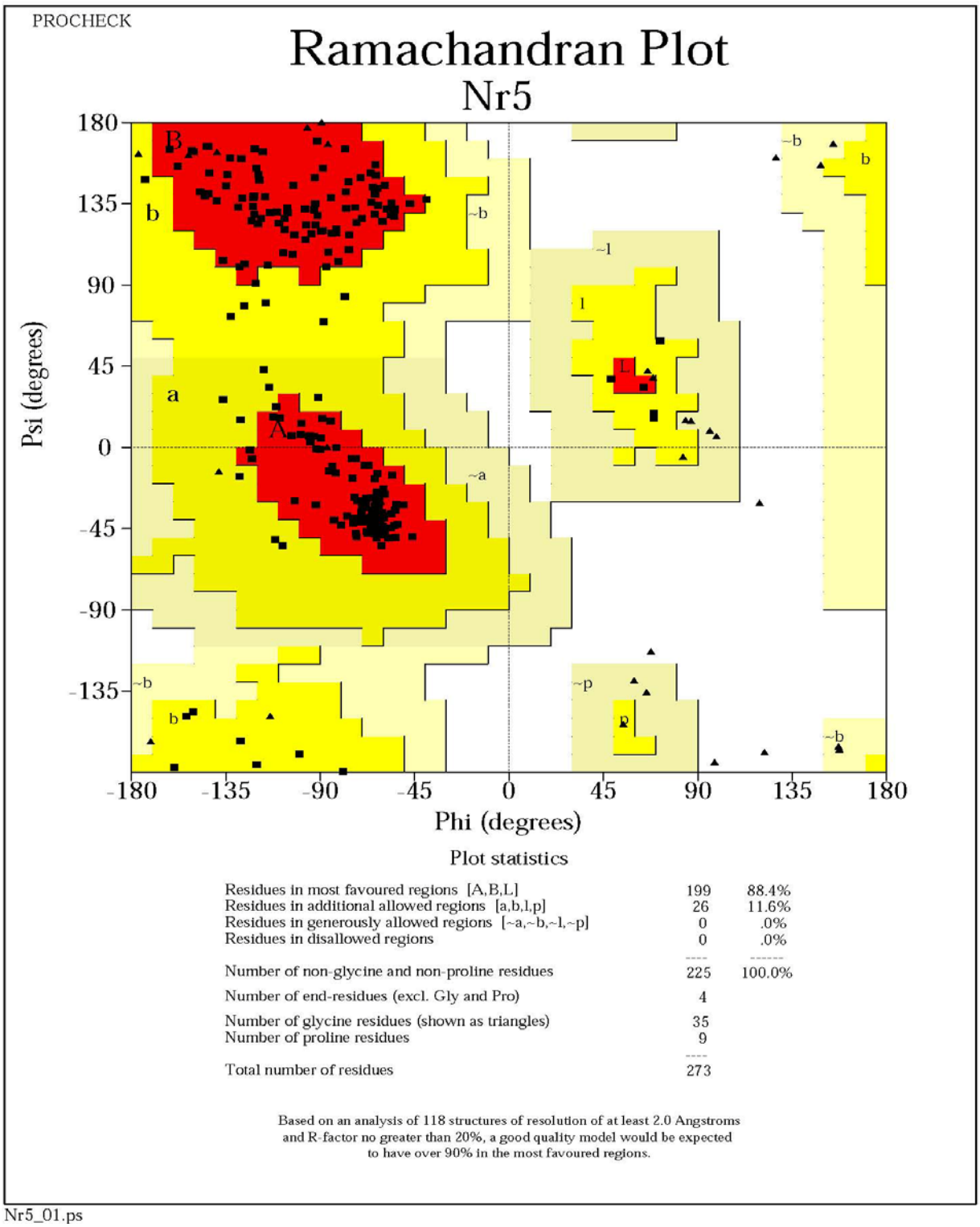
GENE ID: 5620238 aprE1 | subtilisin [Bacillus pumilus SAFR-032]
(10 or fewer PubMed links)

Score = 78.2 bits (191), Expect = 3e-13, Method: Compositional matrix adjust.
Identities = 38/46 (82%), Positives = 39/46 (84%), Gaps = 0/46 (0%)

```
PGSc  92   SKFEWAVANNKDVINMSLGGPSSSRALKNAVDTANNRGVVVCAARG  137
           S  EWAVANN DVINMSLGGP+ S ALKNAVDTANNRGVVV AA G
Subt  215  SGIEWAVANNMDVINMSLGGPNGSTALKNAVDTANNRGVVVVAAAG  260
```

*Subt: subtilisin

Appendix V
Procheck results



Plot statistics

Residues in most favoured regions [A,B,L]	199	88.4%
Residues in additional allowed regions [a,b,l,p]	26	11.6%
Residues in generously allowed regions [~a,~b,~l,~p]	0	.0%
Residues in disallowed regions	0	.0%
Number of non-glycine and non-proline residues	225	100.0%
Number of end-residues (excl. Gly and Pro)	4	
Number of glycine residues (shown as triangles)	35	
Number of proline residues	9	
Total number of residues	273	

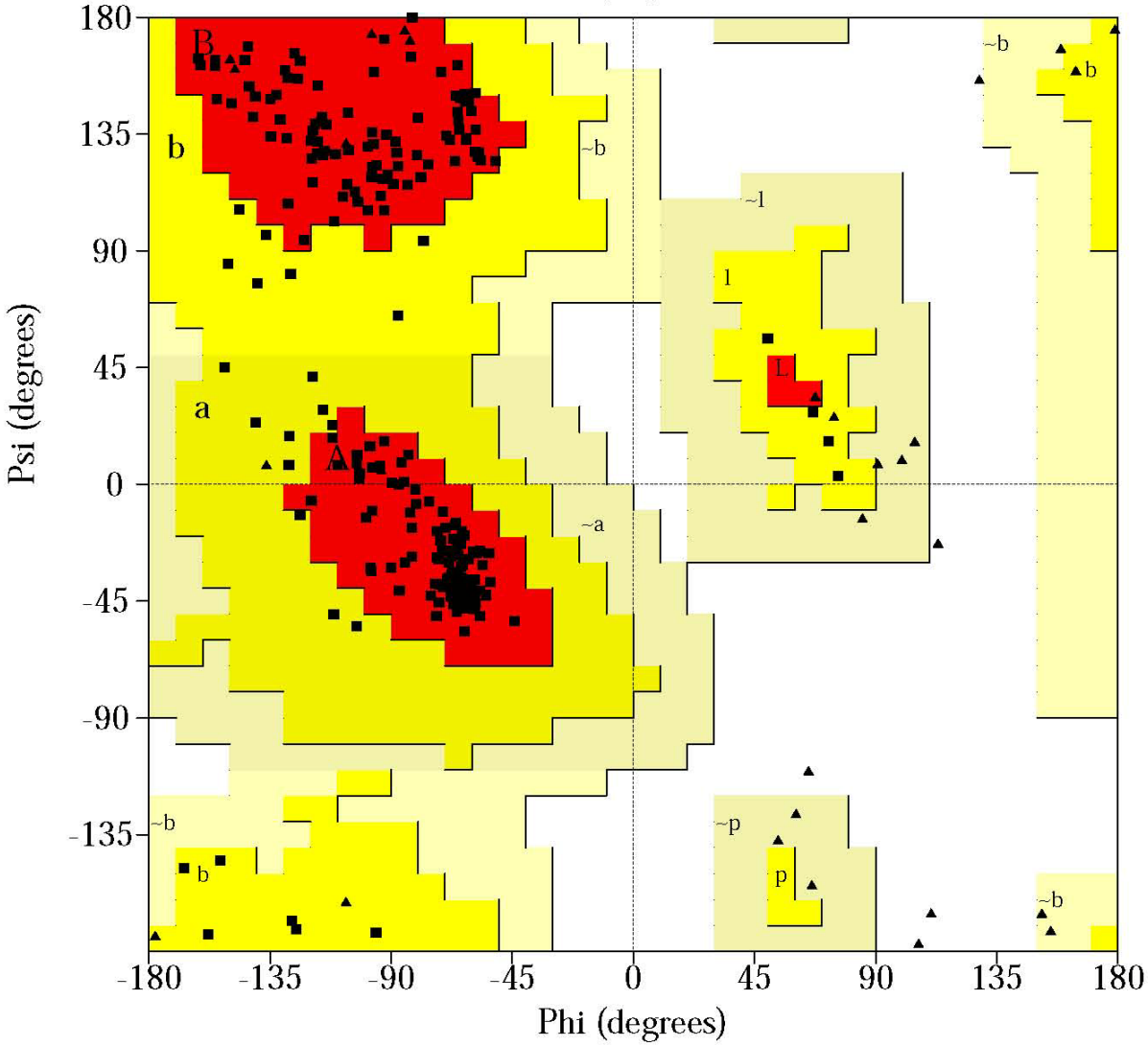
Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Nr5_01.ps

PROCHECK

Ramachandran Plot

Nr8



Plot statistics

Residues in most favoured regions [A,B,L]	201	87.8%
Residues in additional allowed regions [a,b,l,p]	28	12.2%
Residues in generously allowed regions [~a,~b,~l,~p]	0	.0%
Residues in disallowed regions	0	.0%
Number of non-glycine and non-proline residues	229	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	34	
Number of proline residues	9	
Total number of residues	274	

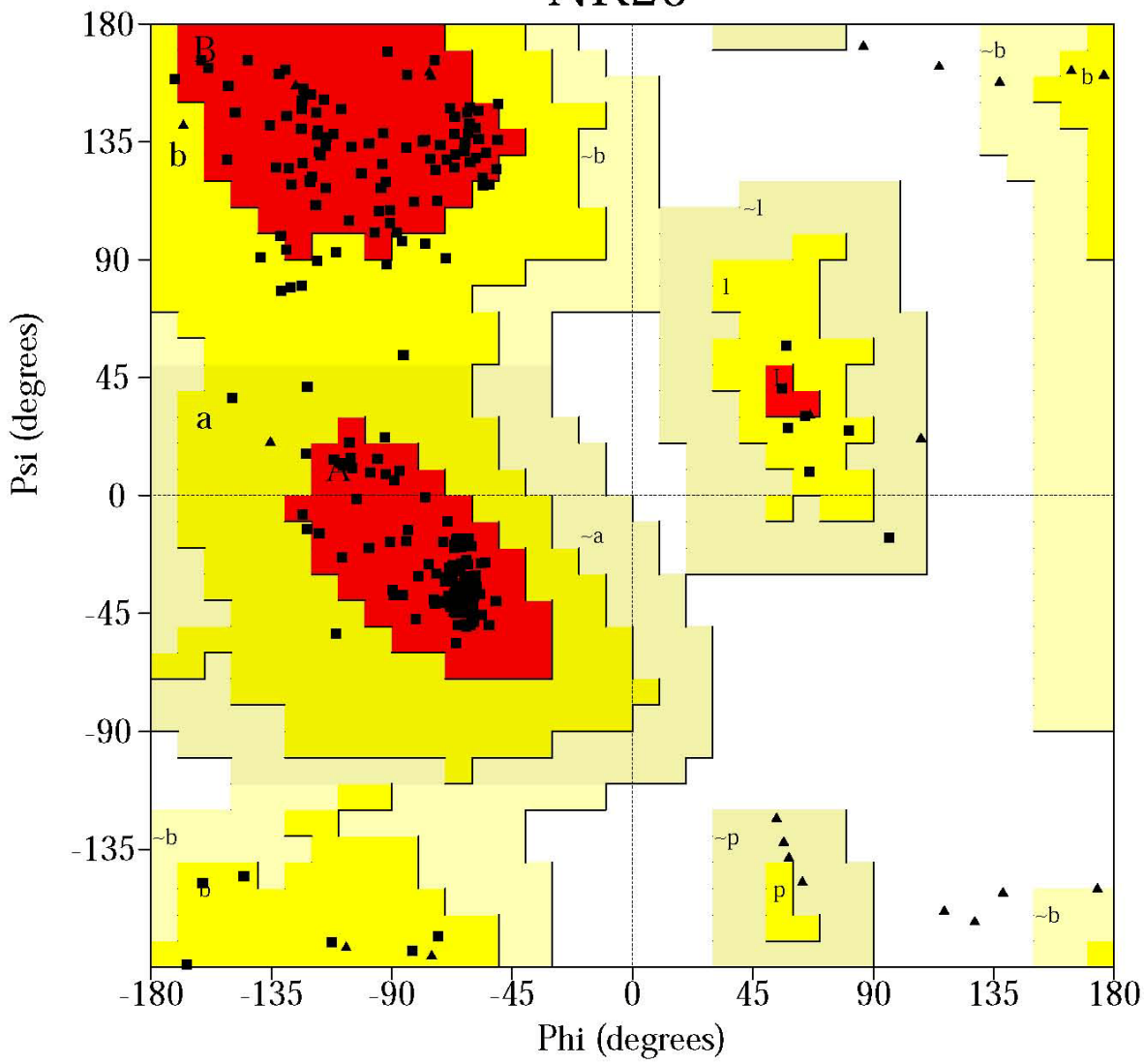
Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Nr8_01.ps

PROCHECK

Ramachandran Plot

NR20



Plot statistics

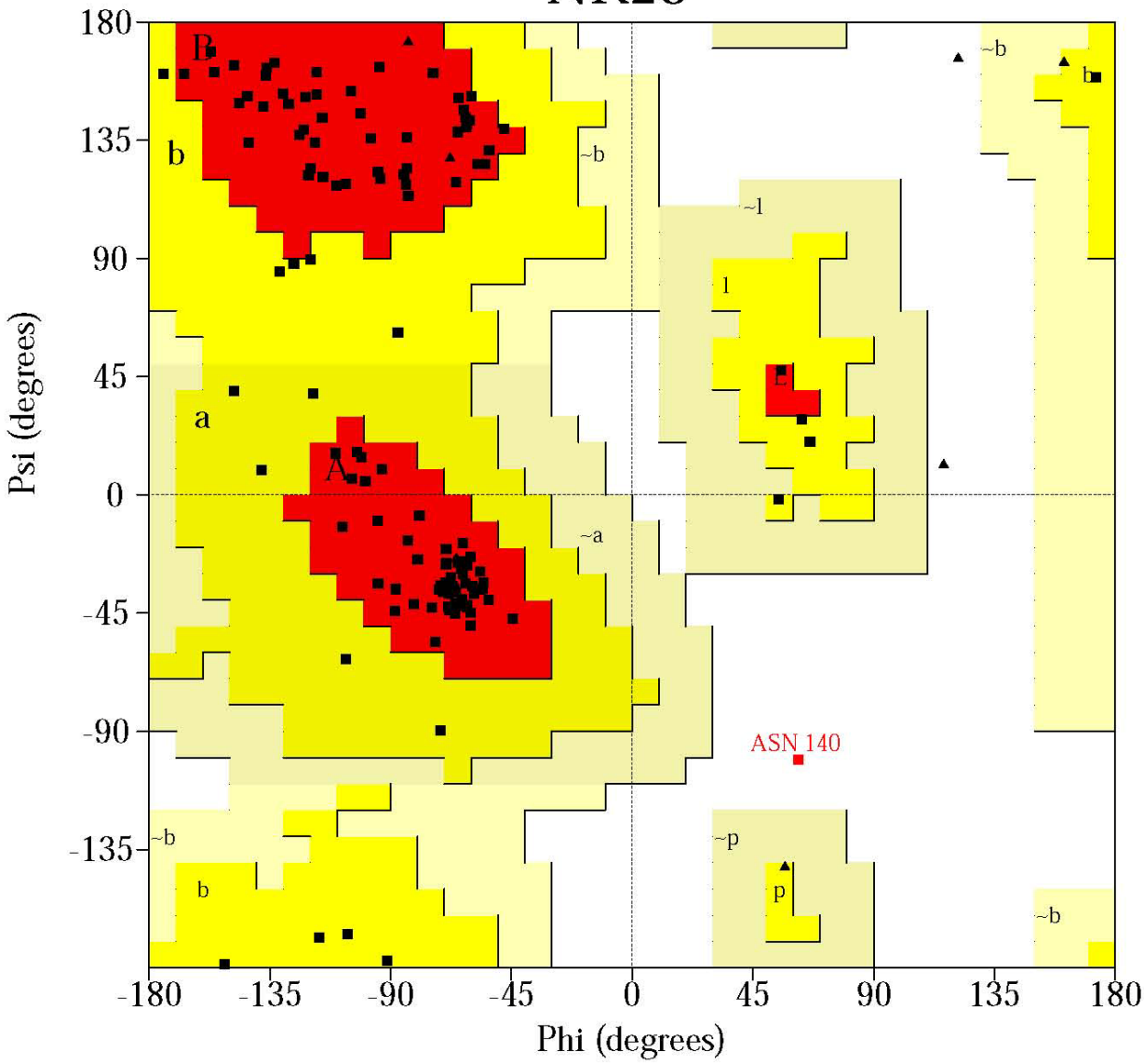
Residues in most favoured regions [A,B,L]	181	85.4%
Residues in additional allowed regions [a,b,l,p]	30	14.2%
Residues in generously allowed regions [~a,~b,~l,~p]	1	.5%
Residues in disallowed regions	0	.0%
Number of non-glycine and non-proline residues	212	100.0%
Number of end-residues (excl. Gly and Pro)	1	
Number of glycine residues (shown as triangles)	27	
Number of proline residues	12	
Total number of residues	252	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

PROCHECK

Ramachandran Plot

NR23



Plot statistics

Residues in most favoured regions [A,B,L]	102	83.6%
Residues in additional allowed regions [a,b,l,p]	19	15.6%
Residues in generously allowed regions [~a,~b,~l,~p]	0	.0%
Residues in disallowed regions	1	.8%
Number of non-glycine and non-proline residues	122	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	9	
Number of proline residues	8	
Total number of residues	141	

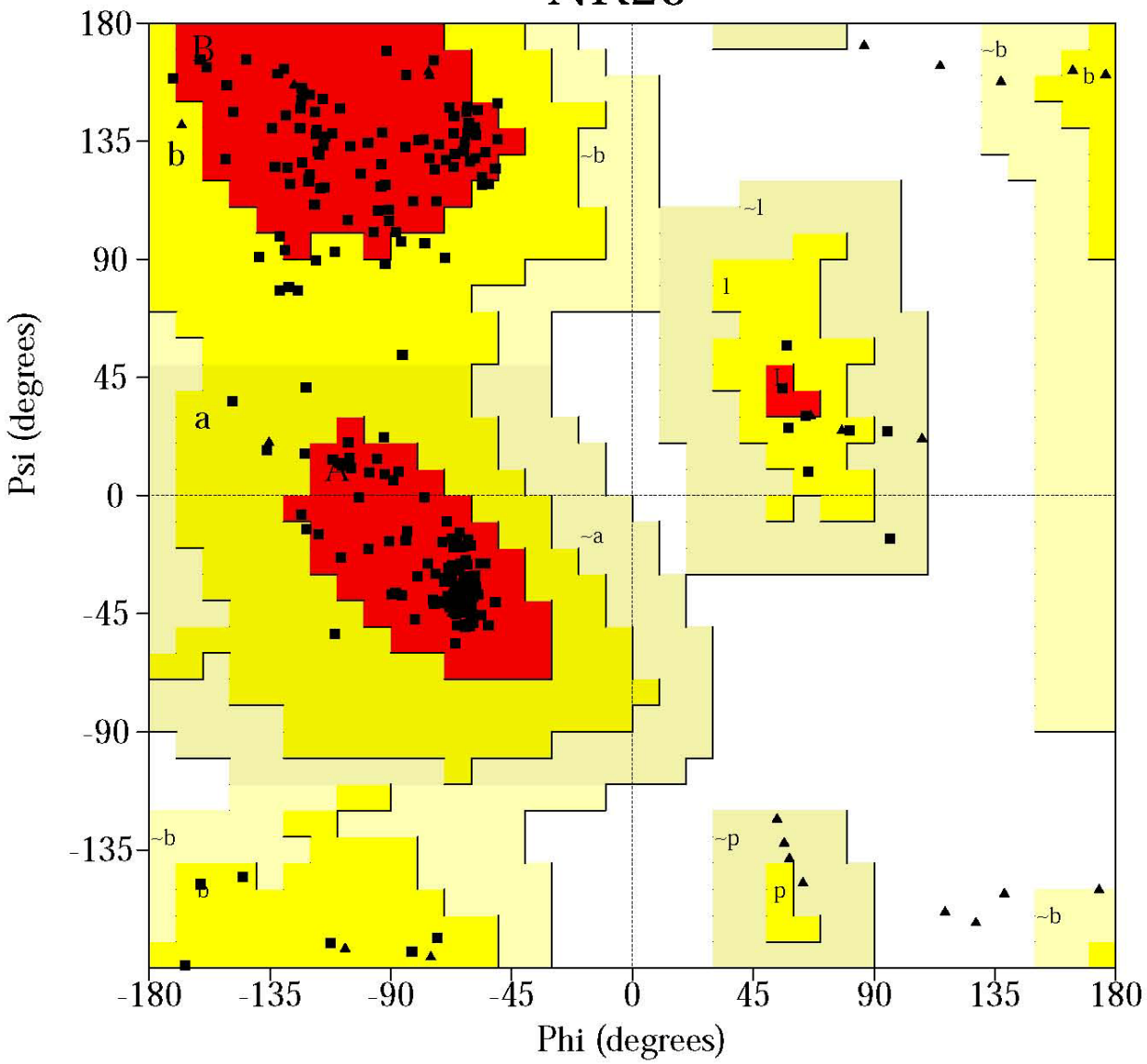
Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

NR23_01.ps

PROCHECK

Ramachandran Plot

NR25



Plot statistics

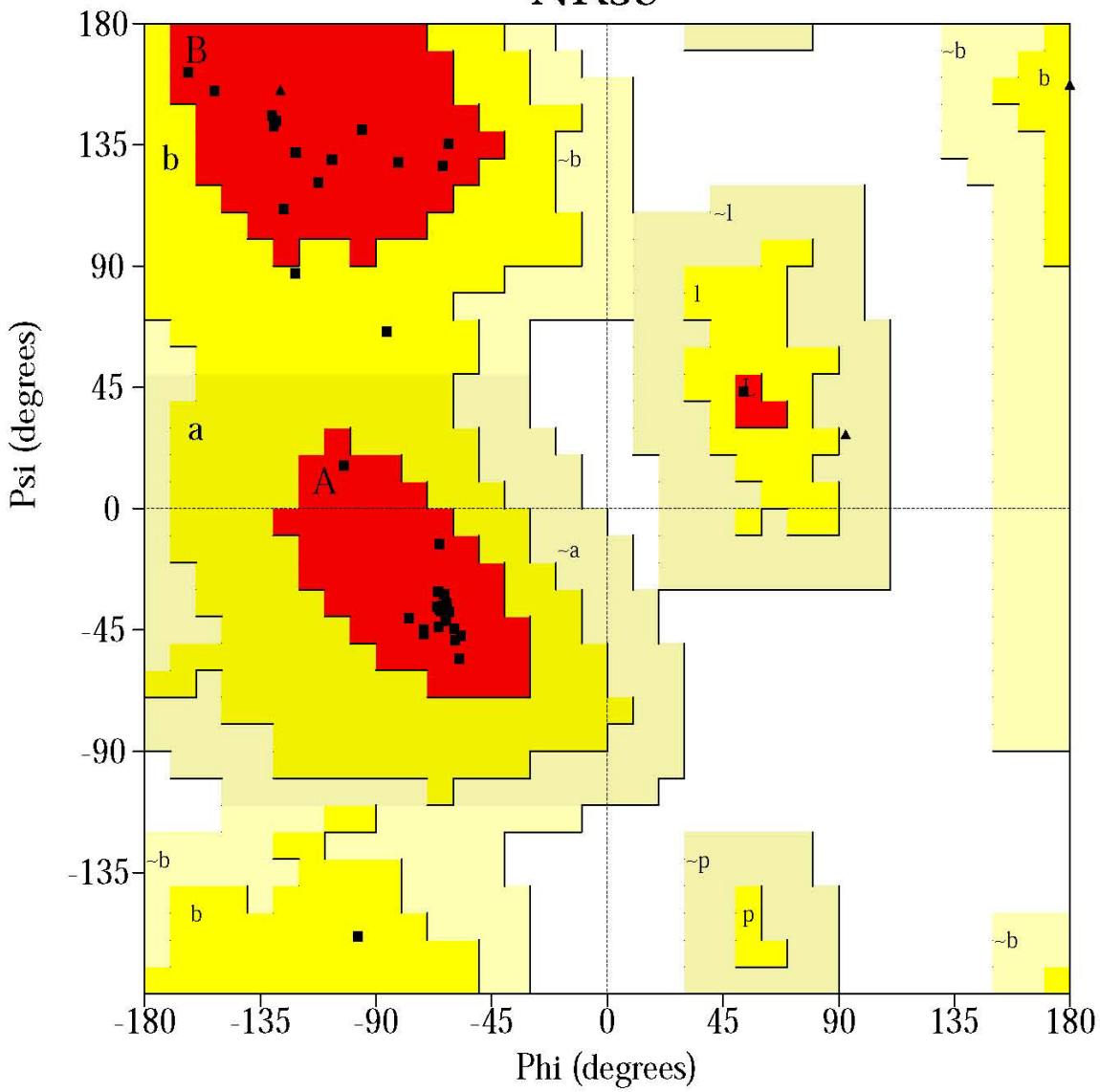
Residues in most favoured regions [A,B,L]	183	84.7%
Residues in additional allowed regions [a,b,l,p]	31	14.4%
Residues in generously allowed regions [~a,~b,~l,~p]	2	.9%
Residues in disallowed regions	0	.0%
Number of non-glycine and non-proline residues	216	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	27	
Number of proline residues	12	
Total number of residues	257	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

PROCHECK

Ramachandran Plot

NRsc



Plot statistics

Residues in most favoured regions [A,B,L]	34	91.9%
Residues in additional allowed regions [a,b,l,p]	3	8.1%
Residues in generously allowed regions [~a,~b,~l,~p]	0	.0%
Residues in disallowed regions	0	.0%
Number of non-glycine and non-proline residues	37	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	3	
Number of proline residues	1	
Total number of residues	43	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.