Investigation on the nature of molecular changes between thermophilic and psychrophilic bacteria: implication on their thermal adaptation

A Thesis

Submitted for the Degree of

DOCTOR OF PHILOSOPHY

By

M.SUBHASHINI
08LPPH28

DEPARTMENT OF PLANT SCIENCES

SCHOOL OF LIFE SCIENCES

UNIVERSITY OF HYDERABAD

HYDERABAD 500 046

INDIA

May 2014

School Of Life Sciences,
University Of Hyderabad,
Central University P. O.,
Hyderabad 500 046.
India

DECLARATION

I hereby declare that the matter embodied in this thesis is the result of investigations carried out by me in the Department of Plant Sciences, School of Life Sciences, University of Hyderabad, Hyderabad, under the supervision of **Dr. J. S.S. Prakash**.

In keeping with the general practice of reporting scientific observations, due acknowledgement has been made wherever the work described is based on the findings of other investigators.

Date: M. Subhashini

Place:

School Of Life Sciences,
University Of Hyderabad,
Central University P. O.,
Hyderabad 500 046.
India

CERTIFICATE

Certified that the work embodied in this thesis entitled “**Investigation on the nature of molecular changes between thermophilic and psychrophilic bacteria: implication on their thermal adaptation**” has been carried out by Mrs. M. SUBHASHINI, under my supervision and the same has not been submitted elsewhere for a Degree.

Head
(Department of Plant Sciences)

Dr. J. S.S. Prakash
(Thesis Supervisor)

Dean
(School of Life Sciences)

CONTENTS

Chapter 3:

Chapter 4:

Chapter 5:

Dedicated to….

~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~

To my parents
and
my Aunty

~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~

Acknowledgements

It is my pleasure to express my sincere gratitude to my research supervisor **Dr. J. S. S. Prakash**, for providing me with an opportunity to work in his lab. I also thank him for his constant cooperation, encouragement and guidance throughout my work.

I thank the present and former Heads of the Department of Plant Sciences, **Prof. Ch. Venkata Ramana** and **Prof. A. R. Reddy** for allowing me to use the departmental facilities. I thank **Prof. A. S. Raghavendra**, Dean, School of Life Sciences, and former Deans, **Prof. R. P. Sharma** and **Prof. M. Ramanadham**, for allowing me to use the general facilities of the school.

I thank my Doctoral Committee members **Prof. P. B. Kirti** and **Prof. P. Appa Rao** for their valuable suggestions.

My special thanks to **Prof. P. B. Kirti, Prof. Lalitha Guruprasad and Dr. Prakash prabhu** for their valuable suggestions and providing their lab facilities.

I would like to thank all my colleagues Dr. K. Sirisha, Bhavani, Dr. Sankar, Arun, Radha rani, Rithika, Deepak, Sanjay, Bantu, Afshan, Ranjith, Harish, Vignan, Dr. Maha lakshmi, Malathi, Triveni, Joystna, Akansha, Dr. Pawan, Naveen, and Dr. Dilip for their help and cooperation and creating a pleasant working atmosphere in the lab. I would like to thank my project juniors Kshitjit and Sumeet for scripting program.

I thank all the faculty members of the School of Life Sciences for their help and allowing me to use the facilities. I thank all the research scholars of the School of Life Sciences for their support. All the non-teaching staff of the School is acknowledged for their help. I also thank Ramesh and Sridhar our lab attenders for their help.

I acknowledge the CSIR for financial aid (fellowship in the form of JRF and DSRF). DST-FIST, UGC-SAP, DBT-CREEB, are also acknowledged for school and departmental facilities.

I am wordless to express my gratitude to my Mother, Father and my Aunty (R. Deevanarthi) who have been the source of encouragement and guiding spirit and for their unfailing love all these years. I owe a great deal of loving thanks to my family Srinivasa rao, Dr. Sai Ramani, Sailaja, Hemanth kumar (Bavagaru), Murari and their kids Rishi, Lohith, Sipivisht and Khaniskaa for their support through out my career.

I take this opportunity to extend my sincere thanks to my husband Dr. Vinay kumar without his support I couldn't have completed my PhD successfully. I also thank my family my son Ruthvik (golu), Varsha, Juhi, Bittu, puneeth, Swathi, Prashanthi, Ashwini, Sri Nagesh, my father in law and others for their love and well wishes.

Subhashini

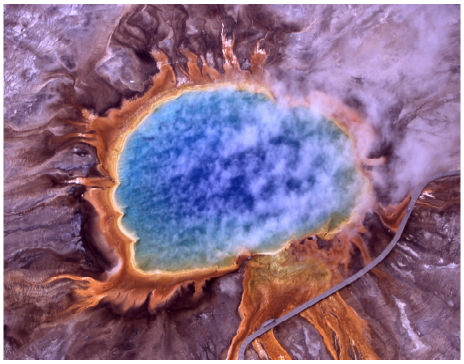
Abbreviations

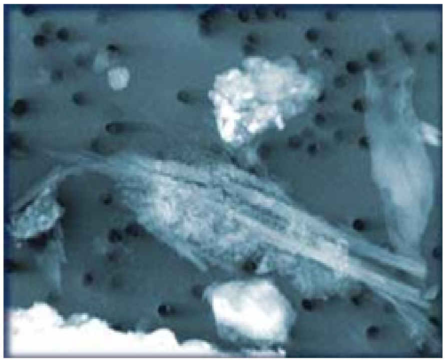
DNA	Deoxyribonucleic acid
°C	Celsius
GC	Guanine Cytosine
AT	Adenine Thymine
SYN	Synonymous
NCNS	Non-conserved Non-synonymous Substitution
CNS	Conserved Non-synonymous Substitution
CIRCE	Controlling the Inverted Repeat of Chaperone Expression
HrcA	Heat shock regulation at CIRCE elements
HSP	Heat Shock Proteins
mRNA	Messenger Ribonucleic acid
M	Molar
NaCl	Sodium Chloride
Atm	Atmosphere
RNA	Ribonucleic acid
3D	three-dimensional
T_{opt}	optimum growth temperature
Min	Minutes
IR	inverted repeat
kDa	Kilo Daltans
A^o	Angstrom
ATP	Adenosine triphosphate
ADP	Adenosine diphosphate
Mg/ml	milligrams/milliliters
NCBI	National Center for Biotechnology Information
BLAST	Basic Local Alignment Search Tool
BDBH	Bi-directional best-hit methods

E-value	Expect Value
MEGA	Molecular Evolutionary Genetic Analysis
SNAP	Synonymous And Non Synonymous Analysis Program
BLOSUM	BLOcks SUbstitution Matrix
MEME	Multiple Expectation maximum for Motif Elicitation
ORF	Open Reading Frame
Ni-NTA	Nichel nitrilotriacetic resin
HIS tagged	Histidine tagged
LB	Luria Broth
µg/µl	microgram/ microliter
IPTG	Isopropyl β-D-1-thiogalactopyranoside
Tris-HCL	2-Amino-2-(hydroxymethyl)-1,3-propanediol hydrochloride
mM	millimolar
NaH_2PO_4	Monosodium Phosphate
µM	micromolar
SDS	Sodium Dodecyl Sulfate
nm	nanometer
Mb	mega bytes
SD	Standard Deviation
A, T, G, C	Adenine Thymine Guanine Cytocine
T_m	Melting Temperature
tRNA	Transfer Ribonucleic Acid
BPROM	Bacterial promoter

List of figures

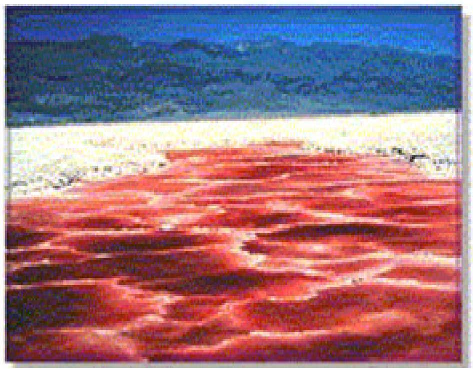
List of tables





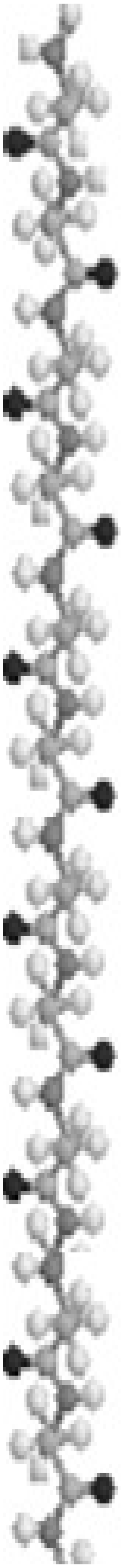
Chapter 1

Introduction



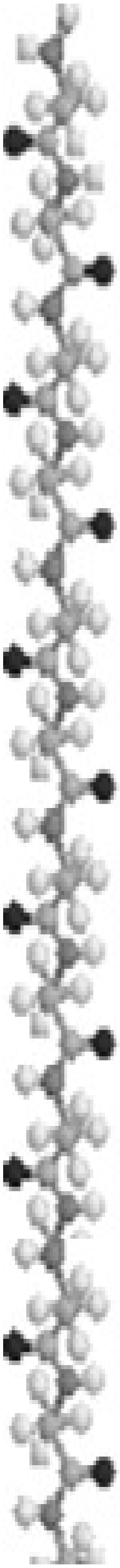
Chapter 2

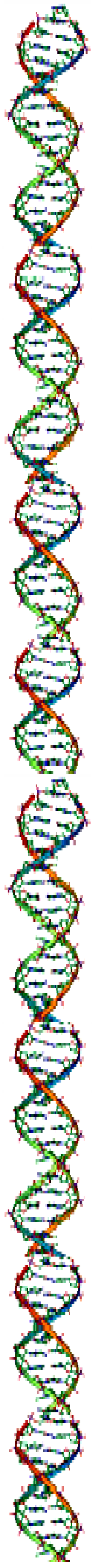
Methodology



Chapter 3

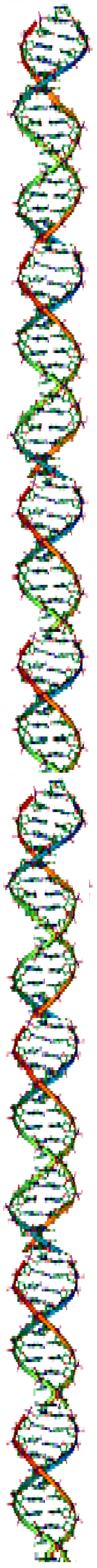
Comparative analysis of proteins between thermophiles and psychrophiles





Chapter 4

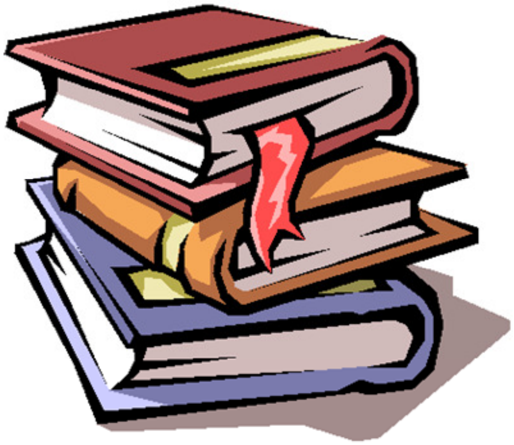
Codon usage differences and rate of synonymous and non-synonymous substitution between thermophilic and psychrophilic conserved orthologs



Chapter 5

Comparison of transcriptional regulation of *groESL* between thermophiles and psychrophiles

References



Abstract:

In the present study, we investigated the differences in the amino acid composition, codon usage bias and, nucleotide and amino acid substitutions if any between thermophilic and psychrophilic bacteria to understand the habitat dependent molecular evolution. We selected 55 thermophiles and 12 psychrophiles to carry out the study, whose genomes were sequenced and the data is publically available. Total 119 proteins are found to be conserved and are commonly present in all 55 thermophilic and 12 psychrophilic bacteria. To identify the nature of molecular changes (at DNA and protein level) between thermophilic and psychrophilic bacteria, these 119 proteins and their corresponding open reading frames were used, because they are commonly present (conserved orthologs) in all the selected bacterial species. Amino acid composition of conserved orthologs between thermophiles and psychrophiles showed that the charged amino acids (aspartic acid (Asp), glutamic acid (Glu), lysine (Lys), arginine (Arg)) and aromatic amino acids (phenylalanine (Phe), tyrosine (Tyr)) were relatively more in thermophiles. While polar uncharged amino acids (serine (Ser), threonine (Thr), glutamine (Gln), asparagine (Asn), cysteine (Cys), methionine (Met)) were more in psychrophiles. We have not observed any significant difference in the frequency of occurrence of aliphatic amino acids between thermophilic and psychrophilic conserved proteins, except, glycine which was seen

relatively more in psychrophiles. As an example, two conserved proteins, GroES and GroEL (here after GroESL) orthologs of thermophiles and psychrophiles were compared to identify the amino acid substitutions which are conserved only in one extremophilic group, but not in the other. Phylogenetic analysis of all conserved proteins in selected thermophilic and psychrophilic bacteria, showed two different clusters, one corresponding to psychrophiles and the second corresponding to thermophiles. Thermophilic cluster and psychrophilic cluster were formed probably due to the conservation of certain amino acids within each group of extremophiles. Amino acids and/or motifs necessary for refolding activity of GroESLs were conserved in all orthologs, irrespective of whether they belong to thermophiles or psychrophiles. Apart from these conserved amino acids, specific differences between the two clusters of GroESL protein sequences were noticed. These specific differences between thermophilic and psychrophilic GroESLs are conserved within each cluster, supporting the argument that during the course of adaptation to extreme conditions, selective changes in genetic material led to changes in amino acids for optimized protein function. Such specific differences in amino acids might have arisen during bacterial adaptation to extreme temperatures. The specific amino acid differences, nature of amino acids and their interactions are critical for GroESL temperature optima as evidenced from the comparative analysis performed in this study. We generated a consensus psychro-GroESL and thermo-GroESL representing psychrophilic and thermophilic bacteria, based on frequency of occurrence of amino acids at each position in a multiple sequence alignment of GroESL orthologs of psychrophiles and thermophiles. These consensus GroESL proteins were reverse translated using the *E. coli* genetic code and the synthetic psychro-*groESL* and thermo-*groESL* were generated. These synthetic genes were cloned in pET28a(+) expression vector, expressed and purified for thermal denaturation studies. The thermal denaturation studies indicated that the thermo-GroEL is more stable than psychro-GroEL. Thermal denaturation studies showed that up to nearly 55 °C the thermo synthetic GroEL protein retains 10 % of its secondary structure where as the psychro synthetic GroEL protein loses approximately one third of its structure at the same temperature. It is well known that genomes of thermophilic bacteria have relatively high GC content when compared to psychrophilic genomes. Thermophiles maintain high GC content in their genomes to stabilize DNA and to prevent the thermal denaturation of DNA at high temperatures. The average GC content of thermophilic and psychrophilic genomes are 53.36 % and 41.83 % respectively. We hypothesized that thermophiles might have evolved with high GC content in their genomes, probably by biased usage of codons. Codon usage analysis of conserved proteins showed that the codons ending with A/T were preferred mostly by psychrophiles, where as codons ending with G/C were preferred by thermophiles. We investigated the nature of nucleotide substitutions, which might have occurred and accepted during the course of adaptation to either temperature extremes. Substitution rate analysis showed d_{SYN} (number of SYNonymous substitutions over synonymous sites) were relatively high in psychrophiles (0.650 ± 0.158), while d_{NCNS} (number of Non-Conserved Non-Synonymous substitutions over sites) more in thermophiles (0.294 ± 0.034). High temperature adaptation favors relatively high rate of non-synonymous substitutions, suggests that bacteria evolved as thermophiles under positive selection pressure. Thermophiles experience more constraints on the genome than psychrophiles. As a result they might have undergone the non-conserved substitutions, which are not naturally acceptable and might have carried forward for adaptation to high temperature. As d_{CNS} (number of Conserved Non-Synonymous sites over conserved non-synonymous sites) was equally distributed between extremophiles, during course of evolution, conservative substitutions have a greater chance of being acceptable. We have observed that the mechanism of *groESL* gene regulation is different between thermophiles and psychrophiles. In most of the thermophiles, majority of heat shock genes and operons are found to be regulated by CIRCE / HrcA system. However, in the genome of psychrophiles, CIRCE element and *hrcA* gene were not found. Instead, RpoH binding site was found in the upstreams of *groE* and *dnaK* operons in psychrophiles. Absence of CIRCE element and *hrcA* gene in psychrophiles infers that negative regulation of heat shock genes might not be necessary as they are rarely exposed to heat. They probably need an immediate activation of *hsp* gene expression upon heat shock. However, thermophiles experience high temperatures at all the time, therefore, they might have evolved with negative regulatory mechanism to maintain the mRNA and protein levels of heat shock proteins in the cell. Overall, the data generated, explains the thermal adaptation strategies used by either extremophiles and provides insight into molecular evolution of thermophiles and psychrophiles to either extreme temperatures.

Chapter 1

I. Introduction

1.1 Extremophiles: Life in extreme conditions:

An extremophile is a living organism that can survive in extreme environmental conditions. Extreme conditions includes hot niches, ice, salt solutions, acid and alkaline conditions, toxic waste, organic solvents, heavy metals, or several other habitats that were considered inhospitable for life (Rampelotto 2013). Extremophiles are found in all the three domains eucarya, bacteria and archaea, while most of extremophiles are predominant in archaea. Extremophiles are divided into two categories: extremophilic organisms that grow in extreme conditions and extremotolerant organisms that tolerate extreme condition and also grow optimally at normal conditions (Rampelotto 2013).

Microorganisms are classified into three groups, psychrophiles, mesophiles and thermophiles based on their growth temperature. Psychrophiles grows in cold temperatures, between -20 °C to +10 °C; mesophiles grows best at moderate temperature ranging from 20 °C to 45 °C and thermophiles grow at very high temperature ranging between 45 °C and 122 °C (Willey et al. 2008; Madigan and Martino 2006; Takai et al. 2008) (Fig 1). Bacteria that are adapted to survive under extremely high and low temperatures are called thermophiles and psychrophiles respectively.

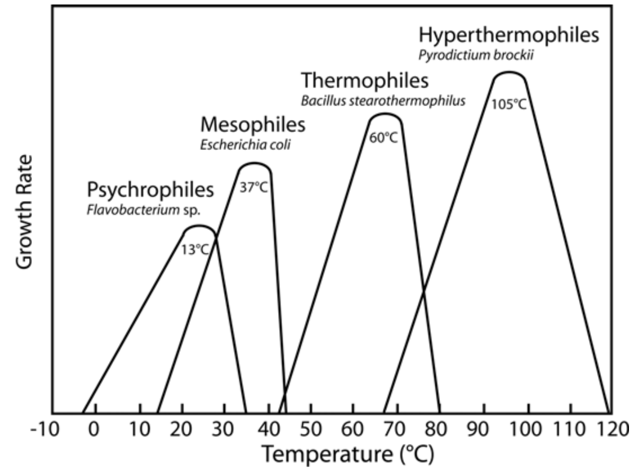


Fig 1: Relation between temperature and growth rate for a typical psychrophilic, mesophilic, thermophilic and hyperthermophilic microorganism. The respective optimal growth temperatures T_{opt} are indicated on the graph (Taken from Madigan et al. 1997).

Bacteria are also found in extremely acidic and alkaline pH environmental conditions and are called as acidophiles (≥ pH 2.0) and alkaliphiles (pH 8.5 - 11) repectively. Some bacterial species existing in high salt conditions are referred as halophiles (2M - 5M NaCl) while those existing in high pressure conditions are refered as barophiles (>1000 atm). Most of the acidophiles fall under archaea domain, which includes *P. torridus* and *P. oshimae*, metabolize optimally at pH 0.7 and grows up to 65 °C, while *Sulfolobus* and *Acidanus* spp. grow up to 96 °C at pH 1–5 (Baker-Austin and Dopson 2007). Other acidophilic archaea includes

Thermoplasma and *Ferroplasma* which metabolize at pH 1-5 (Baker-Austin and Dopson 2007). Bacteria are less tolerant to extreme acidity, but *T. aotearoensis* tolerates pH as low as 3.8, grows up to 60 °C (Hogan 2012). Different kinds of alkaliphiles, including bacteria belonging to the genera *Bacillus*, *Micrococcus*, *Pseudomonas* and *Streptomyces* and eukaryotes such as yeasts and filamentous fungi, have been isolated from a variety of environments (Duckworth et al. 1996; Groth et al. 1997; Horikoshi 1991). Halophiles mostly distributed in archaea domain survive in hyper-saline conditions such as salt lakes, rock salt etc. Strains of *H. subglaciescola*, showed salt tolerance from 0.5 to 20 % and were able to grow at temperatures below 0 °C (Kunte et al. 2002). Apart from this, mixed extremophiles are also reported which survive in two or more extreme conditions like, *C. paradoxum*, a novel alkali-thermophile, which survives in pH range between 7.6 to 9.8 (Cook et al. 1996). *T. alcaliphilus* sp. hyperthermophilic archaeon grows at temperature between 56 to 90 °C and pH between 6.5 - 10.5 (Tanabe et al. 1988).

Almost all extremophiles were isolated and studied extensively over the last two decades, but most of the research was done on thermophiles as they form the source of many novel enzymes that are extensively used in almost all fields of biotechnology (Niehaus et al. 1999). As enzymes from extremophilic organisms are relatively stable and active at extreme conditions, they are preferred over normal enzymes in many enzyme based industries (Niehaus et al. 1999). Therefore, studies to understand the mechanisms of adaptation to extreme conditions by extremophilic organisms and characterization of extremophilic enzymes gained significant attention by the researchers. Moreover, it is important to identify the molecular changes, in comparison with normal enzymes, that might have led an extremozyme to be able to optimally perform at extreme conditions.

1.1.1 Psychrophiles: life at low temperature:

Psychrophiles are the microorganisms that live in cold environments, such as polar regions, the deep sea and cold currents in the ocean. Word “psychro” means cold and “philos” means loving i.e., cold loving. Psychrophiles are also called as “Cryophiles” and “Rhigophiles”. Seventy one percent of the earth surface is covered by oceans with an average yearly temperature of 5 °C (Rodrigues and Tiedje 2008). The extreme cold environments like oceans which have an average yearly temperature of 5 °C, and the ocean depths which have a constant temperature between 1 to 4 °C throughout the year, are the favorable conditions for the psychrophiles to grow (Feller 2013). Permafrost is defined as soils or sediments continuously exposed to a temperature of 0 ºC or below for at least 2 years (Mueller 1973). Permafrost temperatures range from -10 ºC to -20 ºC in the arctic to -10 ºC to -65 ºC in the antarctic, with low water activity and 0.85 – 1 % of carbon. A variety of psychrophilic microorganisms isolated from the kolyma permafrost of northeast Siberia, includes *Arthrobacter, Exiguobacterium, Flavobacterium, Sphingomonas and Psychrobacter* (Vishnivetskaya et al. 2006). Microorganisms were also found in frozen ice with a very small film of water surrounding the cell that serves as a reserve of substrates (Rohde and Price 2007). In comparison to the frozen ice, permafrost is a more favorable environment as it provides heterogeneous soil particles and larger reservoirs of nutrients. Psychrophiles are also found in brine veins between polar sea ice crystals at -20 ^{o}C, in super-cooled cloud droplets and porous rocks in Antarctic dry valleys at -60 ^{o}C (Deming 2002; Bakermans and Skidmore 2011; Sattler et al. 2001; Vaïtilingom et al. 2010; Friedmann 1982; Cary et al. 2010). Hence understanding bacterial adaptation to cold environment is necessary to identify the mechanisms evolved to maintain active cellular processes.

Psychrophiles are abundantly distributed in all three domains bacteria, archaea and eukarya. Polar fish survive beneath the icepack was the biggest psychrophiles and the glacier ice worms which complete their life cycle in the glacier ice (Eastman 1993; Giordano et al. 2012; Roman Dial et al. 2012). Low temperature adapted microbes include mainly bacteria, fungi (in particular yeasts) and microalgae. Among the bacteria commonly reported Gram-negative microorganism which include α-, β- and γ-proteobacteria (*Pseudomonas* spp. and *Vibrio* spp.) and the

Cytophaga–Flavobacterium–Bacteriodes phylum adapt to low temperature. *Coryneforms*, *Arthrobacter* sp. and *Micrococcus* sp. are the most frequently found Gram-positive bacteria that grow at low temperature conditions (D'Amico et al. 2006). Among cyanobacteria, *Oscillatoria*, *Phormidium* and *Nostoc commune* are dominant in most of the antarctic habitats (Pandey et al. 2004). Psychrophilic yeasts, particularly *Cryptococcus* spp., are dominant in most of the antarctic desert soils (Vishniac and Klinger 1986).

1.1.2 Low temperature challenges:

A psychrophilic bacterium, *M. profunda* shows maximum growth rate at 2 ^{o}C and a maximum growth temperature of 12 ^{o}C (Xu et al. 2003). This suggests that even at low temperature, the psychrophilic enzymes are functional and do not undergo any temperature damage. Low temperature challenges include cytoplasmic membrane permeability, increased water viscosity, decreased diffusion rate, rigidity of the enzymes, protein cold denaturation, improper protein folding, formation of secondary structure of DNA and RNA molecules, low cellular activities such as transcription, translation etc., (D'Amico et al. 2006). How the psychrophiles are coping with low temperature effects, is always a topic of research interest? Low temperature adapted bacteria synthesize unsaturated, polyunsaturated, methyl-branched fatty acids and anteiso-branched fatty acids for maintaining the membrane fluidity (Chintalapati et al. 2004; Russell

1997). Anti freeze proteins and cryo-protectants such as trehalose, exo-polysaccharides play important roles in adaptation to low temperature. Anti freeze protein binds to the ice crystals and lowers the temperature so that organism can grow optimally and cryo-protectant like trehalose prevents protein from denaturation and aggregation (Jia and Davies 2002; Phadtare 2004). Apart from this, cold shock proteins, which are expressed at high levels, are involved in protein folding, maintenance of nucleic acid structure, transcription, translation, and membrane fluidity (Phadtare 2004).

Enzymes from a cold adapted organism have increased structure flexibility even though they share high similarity with the mesophilic and thermophilic homologous proteins. Enzymes from the cold adapted organisms show increased catalytical efficiency and decreased temperature stability when compared to mesophilic and thermophilic enzymes (Collins et al. 2003; D'Amico et al. 2003). Structural and thermodynamic basis of cold adaptive enzymes are difficult to obtain as the 3D crystallographic structure of cold adaptive enzymes in their native conformation are difficult to isolate (Feller 2013). Increased catalytical efficiency of a psychrophilic enzyme may be due to their less affinity for the substrate molecule. So that more number of substrate molecules are converted into product resulting in lowering the free energy of activation (Low et al. 1973). The reduced stability in psychrophilic enzymes was due to the more flexible structure resulting in conformational changes of protein molecule. Stability of a psychrophilic protein is also maintained by various factors which includes the reduction of the number of ion pairs; hydrogen bonds and hydrophobic interactions; increased inter-subunit interactions; increased interaction with the solvent; a reduced apolar fraction in the core; higher accessibility to the active site; increased exposure of apolar residues to the solvent; decreased cofactor binding; cluster of glycine residues and a lower proline and arginine content (Violot et al. 2005).

1.1.3 Thermophiles: life at high temperature:

Thermophiles are those organisms that thrive at high temperatures between 45 to 80 ^{o}C. Microorganisms surviving above 80^{o}C are called extreme thermophiles or hyper-thermophiles, which optimally grow close to the boiling point of water (Stetter 1986). Most of the thermophiles and hyper-thermophiles are isolated from continental mud-pools (*S. solfataricus*), hot springs and geysers (*T. aggregans*), deep sea sediments or vents (*M. kandleri*), heated industrial environments, like the outflows of geothermal power plants (*P. islandicum*) and sewage sludge systems (Vieille and Zeikus 2001). Thermophiles are also being isolated from the ordinary surroundings like thermophilic *bacilli* and *clostridia* from compost piles or *Thermus* species from hot water boilers (Sterner and Liebl 2001). Most of hyper-thermophiles belong to the archaeal domain and only few species belong to bacterial domain. About 70 species of hyper-thermophiles belong to archaea and only few species such as *T. maritima* and *A. pyrophilus* which have T_{opt} in the range of 90 to 95 °C are bacterial hyper-thermophiles (Huber et al. 2000). Some archeal members of the genera *Pyrobaculum, Pyrodictium, Pyrococcus* and *Methanopyrus* show the growth temperature 103 - 110 °C (Stetter 1996). Another archeal member *P. fumarii* show the growth temperature range of 90 to 113 ^{o}C (Vieille and Zeikus 2001). The Crenarchaeon hyper-thermophile "Strain 121", which is closely related to *P. occultum* and *P. aerophilum,* has been isolated, grows at 121^{o}C with doubling time of 24 h (Kashefi et al. 2003). Vegetative culture of *Pyrolobus* and *Pyrodictium* are able to survive even after autoclaving at 121^{o} C for 1 h (Blöchl et al. 1997).

It is an interesting area of research to study on how these exciting microbes synthesize and maintain enzymes at high temperatures and how they have optimized the protein function during the course of their thermal adaptation to enable designing of efficient thermo-stable enzymes that can work at high temperature conditions (Kumar and Nussinov 2001).

1.1.4 High temperature challenges:

Temperature is one of the most important environmental factors influencing bacterial growth and activity. High temperature affects the structure and stability of DNA, protein and other biomolecules (Berg et al. 2002). DNA is susceptible to depurination and cytosine deamination at high temperature thereby inducing DNA damage and subsequent denaturation (Forterre in press). As the temperature rises, rates of chemical and enzymatic reactions in the cell also increase. However, this works only up to certain temperatures above which proteins, nucleic acids and other cellular compounds become irreversibly damaged. High temperature affects the protein structure and function by breaking covalent bonds, ionic and electrostatic interactions in the proteins. Breaking of the non-covalent interactions of protein molecules results in structural deformation or conformational disorders. Hyper-thermophilic enzymes are more rigid than their mesophilic homologues as rigidity is a prerequisite for high thermostability of protein (Vieille and Zeikus 2001). How the thermophilic proteins avoid denaturation at high temperature and why mesophilic and psychrophilic proteins are susceptible to denaturation with an increase in temperature?

Protein stability refers to the preservation of the chemical and three dimensional structure of the polypeptide chain and is maintained by interactions among amino acids distributed throughout the protein (Jaenicke 1991). Major research was done to understand the protein stability of thermophiles considering protein sequences that determine the functional properties of protein. Sequential stability factors include presence of certain amino acid like Ala, Arg and Tyr in in higher content in thermophiles compared to mesophilic homologs (Kumar et al. 2000;

Vieille and Zeikus et al. 2001). Residues Asn, Gln, Met and Cys occur less frequently in thermophiles, because of their sensitivity to chemical deamination or oxidation at high temperatures (Jaenicke 2000; Kumar et al. 2000). An increased occurrence of hydrophobic residues with branched side chains, increased proportion of charged residues, decreased occurrence of thermoliable residues and increased occurrence of proline residues in the loop contribute to a higher thermostability of proteins (Kumar and Nussinov 2001). Structural stability factors include surface loop deletion, increased helical content, increased polar surface area and better hydrogen bonding (Querol et al. 1996; Vogt et al. 1996; Kumar et al. 2000). Other stabilizing features include an increased number of ion-pair networks, aromatic-aromatic interactions and salt bridges in thermostable proteins, compared to their mesophilic counterparts (Aguilar et al. 1997; Lebbink 1999; Yip et al. 1998). Thermodynamic properties, like enthalpy, entropy and heat capacity are also essential protein structure stabilizing factors (Kumar and Nussinov 2001). It is observed that the adaptation to high temperature includes proteins with more number of subtle interactions that minimize the surface energy and the hydration of apolar surface groups, with buried hydrophobic residues to maximize packing of the core residues. All these features are maintained by the interactions that results in energetically most favorable stable protein. One of the factors that maintain the stability of the DNA is hydrogen bonding between the nitrogenous bases. Thermophiles and hyper-thermophiles have higher GC content, as energy required to break the bonds between GC is more than required to break the bonds between AT. Messenger RNA of thermophiles contains high GC content and more purine nucleotides, thus help to prevent aggregation of mRNA molecule (Lao and Forsdyke 2000). RNA of hyper-thermophiles contains high GC content and modified bases, which stabilize the secondary and tertiary structures (Forterre, in press). GC content in the first, second and third position of the triplet codon, play an important role in adaptation to extreme temperature conditions. Basak et al. (2004), observed codons with relatively higher GC content in third position of codon (GC_3) in thermophilic bacterium *A. aeolicus* than meosphilic bacterium *B. subtilus* (Basak et al. 2004). Methylation of the nucleotides provides thermo protection to DNA molecule in thermophiles, because they contain different methylation patterns when compared to the mesophiles (Gorgan 1998). Other factors like histone and non-histone proteins, post transcriptional modifications either alone or in combination help in nucleic acid stability in thermophiles when compared with mesophiles (Trivedi et al. 2005). Microorganisms employ different strategies during adaptation to extreme conditions.

1.1.5 Applications of thermophilic and psychrophilic enzymes:

Many genes from the extremophiles were successfully cloned into mesophiles which are used now in food, pharmaceutical and chemical industries (Niehaus et al. 1999). In many industries, enzyme processing is done at elevated temperatures, as high temperature results in an increased bioavailability and solubility of organic compounds and also reduced risk of contamination. Thus obtained enzymes are found to have high purity than enzyme purified by other methods. Some of the extremophilic enzymes that are purified and used extensively in biotechnology industries includes extracellular polymer degrading enzymes such as amylase, pullullanases, cyclodextrin glycosyl transferase, cellulases, xylanases, chitinases, proteinases and other enzymes such as esterases, glucose isomerases, alcohol dehydrogenases, DNA modifying enzymes etc., (Niehaus et al. 1999).

For example, thermostable α–amylases, which were characterized from *P. woesei, P. furiosus* and *T. profundus* have an optimum enzyme activity at temperatures 100 °C, 100 °C and 80 °C respectively (Koch et al. 1991; Chung et al. 1995; Kwak et al. 1998; Lee et al. 1996). A variety of heat-stable proteases have been identified in hyper-thermophilic archaea belonging to the genera *Desulfurococcus, Sulfolobus, Staphylothermus, Thermococcus, Pyrobaculum* and *Pyrococcus*. Serine protease of *P. furiosus*, called pyrolysin, was found to be highly stable with a half-life of 20 min at 105 °C (Eggen et al. 1990). Serine protease from another thermophilic bacterium *F. pennavorans* which hydrolyze feather keratin, is optimally active at 80 °C and pH 10.0 (Friedrich and Antranikian 1996). DNA polymerase (*Taq* polymerase) from *T. aquaticus,* which catalyses template-directed synthesis of DNA from nucleotide triphosphates, has maximal catalytic activity at 75 to 80 ^{o}C. *Taq* polymerase has captured immense importance in molecular biology research due to its thermo-stable nature (Niehaus et al. 1999). It has replaced DNA polymerase of *E. coli,* which is unstable at high temperatures in comparision to *Taq* which resulted in high yield of amplified product, high specificity and automation of PCR reaction. *Taq* has a half life at 96 ^{o}C for 1.6 h. Other most commonly used DNA polymerases are *Pfu* polymerase from *P. furiosus* and *Vent* from *T. litoralis* also known as *Tli* polymerase. *Taq* polymerase shows significant similarity (38% identity) with the DNA polymerase of *E. coli* (Lawyer et al. 1989). Comparative analysis of protein sequences of DNA polymerase from *E. coli* and *T. aquaticus* might provide the reasons why DNA polymerase from *T. aquaticus* is stable at high temperature and why not from *E. coli*.

Recent researches on the psychrophiles and their enzymes have attracted the attention due to their peculiar relationship between the activity, stability and flexibility. Psychrophiles can be source of valuable enzymes with novel substrate specificity, affinity and high activity even at mild temperature and fast heat inactivation rate (Feller 2013). Biotechnology application of psychrophilic enzymes include high enzyme activity at low concentration thereby reducing the cost of enzyme preparation and as the enzymes are heat liable they can be selectively inactivated after the reaction by moderate increase in temperature (Feller 2013). Enzymes from thermophiles are active at very high temperatures and show novel properties with high specific activity for countless applications in food processing, medicne and other industries. Advantages of thermozymes includes reduced risk of microbial contamination, increased mass transfer, resistance to chemical detergent, improved susceptibility to proteins, highly stable and active at elevated temperature. For example, esterase gene from *P. furiosus* was cloned into *E. coli*, this esterase protein was found to be more thermostable with half life of 50 min at 126 ^{o}C and thermo active (optimum temperature of 100 ^{o}C) esterase known till date (Ikeda and Clark 1998). Because of wide applications in biotechnology many number of genes from psychrophiles and thermophiles were cloned and expressed in mesophiles (Ciaramella et al. 1995).

1.2 Importance of Comparative genomics of extremophiles:

Genome sequence information which is available for a large number of prokaryotic and eukaryotic organisms, provides huge amount of valuable datasets for understanding similarities and differences between genomes of different organisms. Characterization of whole genome is important to understand the relationship between the structural and functional aspects of a living organism (Miller et al. 2004). It has become a challenge for the researchers to annotate huge biological data. Comparative genome analysis of multiple organisms provides answers to several questions related to evolution, adaptation to environmental niche, disease cause, special traits etc. Comparative analysis employing mathematical and statistical algoritms, computational tools were designed to understand the different types of biological data like genomics, transcriptomics, proteomics, protein or metabolite interactions, structural characterization of proteins etc., to understand molecular level differences between the organisms. Genomes of many organisms were sequenced due to the availability of many tools and methodologies. The University of California Santa Cruz (UCSC) genomic bioinformatics has developed many tools for browsing genome data and aligned annotation tracts in a single window (http://genome.ucsc.edu). Comparative genomics mainly aid in the identification of the orthologs genes and conserved regions that help in measuring the evolutionary distances among the multiple organisms. With the advances in the technology and availability of the computational tools it is possible to understand the molecular mechanisms of microbial adaptation to their respective habitant. Comparison of related genomes could reveal phenotypic differences and recent evolutionary events between the genomes. Genomes evolve through various mechanisms including long- and short-range, intra-genomic rearrangements like displacements and inversions, insertion of laterally acquired DNA, gene loss and single-nucleotide polymorphisms (Abbott et al. 2007). Thus obtained genomic alignment was able to provide insight into the conservation of orthologous genes, evolution of organisms and identification of unique genes that might be responsible for the adaptation to different habitant (Scanlan et al. 2009). Many organisms don't survive the extreme conditions while the extremophiles survive optimally to these adverse conditions. Microorganisms belonging to the same or closely related family but adapting to different environmental conditions could be due to rearrangements, insertions or deletions, transversions or translocations that take place in the genomes during the course of evolution (Hao and Golding 2006). For example, Kube et al. (2010), noticed the absence of virulence factors of plant-pathogenic *Erwinia* strains in non-pathogenic *Erwinia*, by comparative genomic approach (Kube et al. 2010).

Organisms surviving in the different environmental conditions synthesize certain proteins, modulate proteins and secrete certain compounds which are essential for survival to that particular habitat (D'Amico et al. 2006). For example, anti freeze proteins, cryo-protectant, cold shock proteins in psychrophiles (D'Amico et al. 2006), heat shock proteins in thermophiles, osmolytes in halophiles, acidic proteins in alkailophiles etc.,. Acidophilic organisms have an efficient mechanism to pump protons out of the intracellular space in order to keep the cytoplasm at or near neutral pH. Comparative analysis of proteins adapted to neutral and low pH could provide insight into molecular level differences by which proteins can achieve acid stability.

Origin and evolution of extremophiles is important to understand phylogeny of extremophiles. Most of the thermophiles and hyper-thermophiles represent the most ancient forms of life now present on earth (Madigan et al*.,* 1997). Thus it becomes necessary and important to understand the adaptation of extremophiles to extreme conditions by comparative sequence analysis. Genomes of extremophilic organisms surviving in different habitats contain some unique genes in addition to the core genes. Unique genes are retained by natural selection process or evolutionary mechanism which might be necessary for their adaptation or to maintain their population (Scanlan et al. 2009). Identifying genes that are unique to an organism compared to another, could explain its adaptation to a respective habitat. Comparative genomics leads us to analyze that during evolution, gene duplication events, gene rearrangements, gene deletions and insertions, gene loss and gene gain which might help in explaining the evolutionary dynamics. Large number of insertions and deletions can be observed in many bacterial species and can be compared among the species (Mirkin et al. 2003; Hao and Golding 2004; Garcia-Vallvé et al. 2000).

Comparative genome analysis between thermophiles and psychrophiles might reveal reasons for their extreme temperature adaptation.

1.3 Molecular chaperons:

Stress conditions or unfavorable environments like high temperature (heat shock), salt, free radicals, heavy metals, ethanol and amino acid analogues disturb the tertiary structure of proteins and cause adverse effects on cellular metabolism (De Maio 1999). Heat shock proteins are essential molecules of the cell that protects the bacteria from adverse effects. The principal heat-shock proteins that have chaperone activity are classified based on their molecular weight includes, Hsp33, Hsp60, Hsp70, Hsp90, Hsp100 and the small heat-shock proteins (sHSPs) (Li and Srivastava 2004). Heat shock proteins are involved in interaction with non-native protein subunits, stabilization of protein-folding intermediates and prevent protein aggregation (Li and Srivastava 2004). Heat shock proteins are most characterized and well studied in *E. coli* (Baldini et al. 1998). In *E. coli* the heat shock genes are positively regulated by the alternative sigma factors σ^{32} encoded by *rpoH* gene and negatively regulated by the products of heat shock genes, *dnaK, dnaJ* and *grpE* (Bukau 1993; Yura et al. 1993).

Molecular chaperones are proteins, which assist the folding of substrate protein to the native state preventing from misfolding and aggregation. Chaperonins are the essential genes required for survival of bacterial cell and deletion of these genes is lethal to the cell. *E. coli* chaperonin is a well-studied system comprises two proteins belonging to one gene family groE; GroEL ("large", Hsp60 or Cnp 60) and GroES ("small", co-chaperonin, Hsp10 or Cnp 10). In *C. crescentus, groES* and *groEL* exists as single operon under the control of either of the two promoters, heat shock promoter (IR element) and σ^{32}-RNA polymerase or constitutive promoter (Baldini et al. 1998). GroE heat shock proteins are expressed in high levels in a cell in response to different environmental stress conditions (Ellis 1996; Parsell et al. 1993). GroEL (Hsp60) mediates proper protein folding, assembly, transport and degradation of proteins within the cell and also prevents the protein aggregation (Craig et al. 1993; Hendrick and Hartl 1993). These genes are present in all eubacteria and eukaryotic cell organelles such as mitochondria and chloroplasts (Goloubinoff et al. 1989). In archaebacteria and eukaryotic cell cytosol, there are no close homologues of Hsp10 or Hsp60 have been identified. Instead, they contain the Tcp-1 (t-complex polypeptide 1) family of proteins (distant homologues of Hsp60) (Valpuesta et al. 2002).

1.3.1 Structural features of GroESL chaperone:

GroEL is an oligomeric protein, namely tetradecamer. Its 14 subunits are organized in two seven-membered rings, stacked to each other. GroEL exists as two heptameric rings stacked back to back, each containing seven identical 57 kDa subunits, with a central cavity ~45 $\overset{o}{A}$ wide that provides an environment for the efficient folding of proteins (Braig et al. 1994, 1995). *E. coli* GroEL subunit has three domains: an equatorial domain containing an ATP/ADP binding site (residues 6-133 and 409-523), an apical domain (residues 191-376) with the binding site for both nonnative proteins and GroES, and an intermediate domain (residues 134-190 and 377-408) that has a hinge region connecting the equatorial and the apical domains (Braig et al. 1994). GroES has a dome-shaped structure containing seven identical 10 kDa subunits assembled as one heptameric ring that binds to one of the GroEL rings (Hunt et al. 1996). Activity of GroEL/ES depends on the structure of the monomer.

1.3.2 Functional features of GroESL chaperon:

Protein folding is essential to cell or else proteins will get accumulated in a conformation, which are energetically stable, but not a native one. Molecular chaperones play an essential role in the cell by binding to non-native proteins and prevents from protein aggregation. Many unfolded proteins are folded to its active conformation by binding to the GroESL chaperone. Many factors affecting the protein folding include non-covalent interactions and secondary structural elements of protein.

For example, if the hydrophobic residues are exposed outside the unfolded protein, it results in a strong tendency to aggregate. Secondary structure elements and their interactions are essential to form an active protein globule, where hydrophobic patches are buried inside, thus stabilizing the tertiary structure and enabling inter-domain interactions (Manakova 2001). On the other hand, the concentration of the unfolded protein is also important for proper folding; the concentration of proteins in cytosol was estimated to be around 300-400 mg/ml (Zimmerman et al. 1991). Any mis-folded species are immediately transferred to proteases of the protein degradation system (Goldberg 2003). Many small proteins are denatured and refolded into the native state without the use of GroESL, but larger proteins require GroESL complex.

All known substrates of GroEL fall in two groups. Proteins from the first group require the complete chaperonin system for refolding. They are called the stringent substrates (Goloubinoff et al. 1989; Rye et al. 1997). In the second group there are proteins, which could reach a native state after only one round of binding to GroEL. They do not need ATP hydrolysis (Hansen and Gafni 1993; Hayer-Hartl et al. 1996). ATP in this case can be replaced by a non-hydrolysable ATP analogue and ADP. The proteins of the second group usually refold themselves spontaneously in the solution with high efficiency. GroEL in the absence of its cofactors forms a tight complex with these substrate proteins. Addition of a nucleotide results in the release of the substrate by lowering the affinity of GroEL to the unfolded protein. GroEL usually increases the yield of refolding remarkably, especially in the presence of cofactors (GroES and ATP).

1.3.3 Conformational switch of GroESL protein:

GroE system is complex machinery, self-regulated by the interplay of several components. GroEL/ES mediated protein folding involves multiple rounds of binding, encapsulation, and release of substrate protein until the native conformation is reached. Unfolded substrate proteins bind to a hydrophobic region on the interior rim of the open cavity of GroEL, and further binding of ATP molecule to one of the seven subunit rings (the *cis* ring), induces a conformational change, includes the upward movement of the intermediate and apical domain, that allows stable association of the separate lid structure, GroES (Xu et al. 1997; Bukau and Horwich 1998; Hartl and Hayer-Hartl 2002; Sigler et al. 1998; Thirumalai and Lorimer 2001). Binding of GroES, induces multiple rounds of folding, returning each time to its original unfolded state, until the native conformation or an intermediate structure committed to reaching the native state is achieved. ATP hydrolysis in the *cis* ring triggers the release of GroES, ADP and the substrate protein from the GroEL, with the subsequent binging of ATP to the *trans* ring (Rye et al. 1997, 1999). Adenine nucleotides are main allosteric regulators of the chaperone activity of GroEL. The binding of ATP by GroEL ring is a highly co-operative process. Binding of ATP or ADP changes the affinity of GroEL with respect to the substrate protein and GroES (Inobe et al. 2008; Gorovits et al. 1997). GroEL is switched from one mode or state to another, each state having completely different affinities to both its substrates (polypeptide to be folded and ATP, as a substrate for ATPase) after binding of ATP. The cooperative binding of the substrate and ATP molecule to the GroEL have been clearly explained by two approaches: the classical approach of Monod-Wyman-Changeuax and the Koshland-Nemethy-Filmer model for sequential transitions. In the first model GroEL subunits in one of the seven-membered ring undergo two transition states. In the first state the unliganded GroEL has low affinity to ATP and high affinity to unfolded protein substrate and the second state is the ATP bound state of GroEL with the opposite features. The second theory describes the negative co-operatively between two rings of GroEL, each ring comprising a unit of the concerted hydrolysis where all seven subunits act simultaneously. Binding of ATP to one ring allows binding of ATP to another ring only at higher concentration of ATP. Wide range of non specific unfolded proteins binds to GroEL through hydrophobic contacts, mainly to the apical domain ofGroEL (Fenton et al. 1994; Houry et al. 1999; Sakikawa et al. 1999; Viitanen et al. 1992). Apart from GroEL and substrate interactions, GroES interacts with *cis* ring of the GroEL and also with the substrate protein (Mayhew et al. 1996; Weissman et al. 1995). These non covalent interactions (hydrophobic, ion-pair, hydrogen-bonding, vanderwaals interactions) within each protein molecule, strength the effectiveness and molecular compactness, there by increasing the stabilization energy of the protein molecules with respect tof temperature change.

1.4 Significance of study:

Researchers have focused on the molecular characterization of the extremophilic enzymes, as these are relatively stable and active at extreme conditions. Many genes from the extremophiles were successfully cloned into mesophiles, which are used in food, pharmaceutical and chemical industries. For example, Cellulase from a thermophilic bacterium *T. maritima* MSB8 and a psychrophilic bacterium *A. alcalophilum* showed a maximal activity at 95 °C and 40 °C respectively (Bronnenmeier et al. 1995; Hayashi et al. 1996). The difference in temperature optimum of cellulase between psychrophilic (*Acremonium*) and thermophilic (*Thermotoga*) bacteria might suggest that functions of most proteins are optimized to a range of temperature in which the microorganism has been adapted to survive. In another example, optimum temperature required for refolding activity of GroESL of *Oleispira*, a psychrophilic bacterium is between 4-12 ^{o}C (Ferrer et al. 2003). While ATPase activity of recombinant *C. thermocellum,* a thermophilic bacterium Cpn60 showed maximum activity between 60 and 80 °C (Cross et al. 1996). It is most likely that the differences in amino acids at various positions of a functionally similar protein from either extremophiles could be responsible for their differential temperature optima. Thus, differences in temperature optima might be due to the differences in the composition of amino acids between the two groups. Therefore, it is necessary to analyze variation in the amino acid composition between the extremophiles that contribute to the stability at high or low temperature. We hypothesize that an enzyme performing similar function in a thermophilic and psychrophilic bacterium may exhibit variation in its amino acid composition. The key aspect of adaptation to any extreme condition is the optimization of the protein function which is necessary for the cellular metabolism. Adaptive strategies developed by microbes to cope with extreme temperatures are genetically encoded and need to be analyzed to understand the mechanisms involved. Comparative genomics helps in surfacing these mechanisms as the favorable subtle changes in most of the bio-molecules were conserved through time, which might be responsible for the evolutionary success of the organism.

Amino acid composition analysis was performed between psychrophilic and mesophilic proteins to understand the contribution of amino acids in relation to adaptation to low temperature (Metpally and Reddy 2009). Thus it becomes necessary and important to understand the adaptation of bacteria to extreme conditions by comparative sequence analysis. Comparative analysis of conserved orthologous proteins and gene sequences between thermophilic and psychrophilic bacteria is necessary as it may reveal differences at the molecular level which might be because of their course of adaptation to extreme conditions. No reports are available till today and it is important to understand the molecular adaptation of these two extremophilic groups.

The main objective of this study is to better understand the thermal adaptation strategies used by either extremophiles and to gain insights into the molecular evolution of thermophiles and psychrophiles to either extreme temperatures.

1.5 Objectives:

I. Comparative analysis of proteins between thermophiles and psychrophiles

1. To identify the proteins those are common (conserved orthologs) in all thermophilic and psychrophilic bacterial genomes.

2. To identify the differences in the amino acid composition between thermophilic and psychrophilic conserved orthologs that might have contributed to the stability of proteins at high or low temperature.

3. To analyze the differences in amino acids at various positions of GroESL from either extremophiles to identify the amino acids that are responsible for their differential temperature optima.

4. Design and synthesis of consensus psychro-*groESL* and thermo-*groESL* genes, expression and purification of corresponding psychro-GroESL and thermo-GroESLs for thermal denaturation studies.

II. Codon usage differences and rate of synonymous and non-synonymous substitution between thermophilic and psychrophilic conserved orthologs

1. To analyze the codon usage between thermophilic and psychrophilic conserved proteins to understand the evolution of thermophiles with high GC content genomes.

2. To analyze the rate of synonymous and non-synonymous substitutions between thermophilic and psychrophilic conserved orthologs to understand the molecular evolution of either extremophiles.

III. Comparison of transcriptional regulation of *groESL* between thermophiles and psychrophiles.

1. To identify the *Cis* acting elements and to compare the corresponding transcriptional regulators of *groESL* operon between thermophiles and psychrophiles to see any differences in the regulatory mechanism to sustain either extremes temperature.

2. Methodology

2.1 *Insilico* analysis:

2.1.1 Sequence and annotation files retrieval - working data set:

We considered 55 thermophiles and 12 psychrophiles in our analysis for which the genome sequence information is available. We developed a program in *Perl Script* to extract nucleotide, protein sequences and corresponding annotation files (*.faa, *.ptt, *.fna, *.gbk files) for 55 thermophiles and 12 psychrophiles from the NCBI (National Center For Biotechnology Information) (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria).

The genomes that are included in our analysis are listed in the Table 1. The few organisms with different strain names are not included in our analysis due to the complexity of the genomic data, to reduce the redundancy and noise in the data. List of the thermophilic and psychrophilic bacteria with the genome size, number of genes and proteins are given in Table 1. We created a separate local protein, nucleotide and intergenic sequence databases for each selected extremophilic species using the *Perl Script*.

Table 1: List of bacterial species considered to compare the nature of molecular changes between thermophiles and psychrophiles.

Extremophiles	**Genome Size (Mb)**	**No of genes**	**No of proteins**	**Gram staining**	**Growth temperature (^{o}C)**
Psychrophiles					
Flavobacterium psychrophilum JIP02/86	2.88	2556	2446	Negative	3-15
Polaribacter MED152	2.96	2676	2635	Negative	16
Colwellia psychrerythraea 34H	5.37	5054	4910	Negative	8
Photobacterium profundum SS9	6.4	5702	5489	Negative	15
Psychrobacter arcticus/ arcticum	2.65	2211	2120	Negative	–10 ~22
Psychrobacter cryohalolentis K5	3.1	2581	2511	Negative	–10
Pseudoalteromonas haloplanktis TAC125	3.85	3620	3485	Negative	4~15
Psychromonas ingrahamii	4.56	3863	3545	Negative	–12 ~10
Desulfotalea psychrophila	3.66	3322	3234	Negative	7
Shewanella halifaxensis HAW-EB4.	5.23	4464	4278	Negative	10
Shewanella sediminis HAW-EB3	5.52	4666	4497	Negative	10
Aliivibrio salmonicida LFI1238	4.66	4421	3911	Negative	12
Synechococcus sp. JA-3-3Ab	2.93	2897	2760	Negative	50-60
Thermobifida fusca YX	3.64	3161	3087	Positive	55
Thermosynechococcus elongates BP-1	2.59	2525	2476	-	55
Geobacillus kaustophilus HTA426	3.59	3653	3539	Positive	65
Symbiobacterium thermophilum IAM 14863	3.57	3454	3337	Positive	60
Aquifex aeolicus VF5	1.59	1611	1560	Negative	96
Hydrogenobaculum	1.56	1694	1629	Negative	58

sp. YO4AAS1					
Sulfurihydrogenibium sp. YO3AOP1	1.84	1829	1721	Negative	68
Thermus thermophiles HB27	2.13	2263	2210	Negative	68
Fervidobacterium nodosum Rt17-B1	1.95	1846	1750	Negative	80
Thermodesulfovibrio yellowstonii DSM 11347	2	2089	2033	Negative	65
Thermosipho africanus TCF52B	2.02	2041	1954	Negative	75
Thermotoga maritima MSB8	1.87	1946	1891	Negative	80
Nitratiruptor sp. SB155-2	1.88	1910	1843	Negative	55
Synechococcus sp. JA-2-3B'a(2-13)	3.05	2942	2862	Negative	50-55
Anoxybacillus flavithermus WK1	2.85	2954	2831	Positive	60
Caldicellulosiruptor saccharolyticus DSM 8903	2.97	2844	2682	Positive	70
Carboxydothermus hydrogenoformans Z-2901	2.4	2707	2620	Positive	78
Clostridium thermocellum ATCC 27405	3.84	3363	3173	Positive	60
Coprothermobacter proteolyticus DSM 5265	1.42	1541	1482	Negative	63
Heliobacterium modesticaldum Ice1	3.08	3137	2999	Negative	50-52
Thermoanaerobacter tengcongensis MB4	2.69	2721	2588	Negative	75
Chloroflexus aggregans DSM 9485	4.68	3904	3731	Negative	55
Chloroflexus aurantiacus J-10-fl	5.26	3990	3853	Negative	60
Roseiflexus castenholzii DSM 13941	5.27	4509	4330	Negative	50
Moorella thermoacetica ATCC 39073	2.63	2613	2463	Positive	55
Thermoanaerobacter pseudethanolicus ATCC 33223	2.38	2358	2239	Positive	65
Dictyoglomus thermophilum H-6-12	1.96	1973	1912	Negative	78

Dictyoglomus turgidum DSM 6724	1.86	1867	1744	Positive	70-80
Thermomonospora curvata DSM 43183	5.64	5061	4890	Positive	50
Ammonifex degensii KC4	2.16	2239	2080	Negative	70
Caldicellulosiruptor bescii DSM 6725	2.93	2836	2666	Positive	75
Caldisericum exile AZM16c01	1.56	1633	1581	-	68
Chlorobium tepidum TLS	2.15	2337	2245	Negative	48
Geobacillus WCH70	3.15	3590	3168	Positive	65
Streptococcus thermophilus CNRZ1066	1.8	2000	1915	Positive	45
Sulfurihydrogenibium azorense Az Fu1	1.64	1768	1723	Negative	68
Thermanaerovibrio acidaminovorans DSM 6589	1.85	1828	1738	Negative	55
Deinococcus geothermalis DSM 11300	3.25	3142	3054	Positive	47
Petrotoga mobilis SJ95	2.17	2015	1898	Negative	60
Natranaerobius thermophilus JW/NM-WN-LF	3.19	3024	2906	Positive	53
Acidimicrobium ferrooxidans DSM 10331	2.16	2089	1964	Positive	48
Acidothermus cellulolyticus 11B	2.44	2217	2157	Positive	58
Methylacidiphilum infernorum V4	2.29	2524	2472	-	60
Thermomicrobium roseum DSM 5159	2.95	2904	2854	Negative	70
Geobacillus thermodenitrificans NG80-2	3.61	3590	3445	Positive	65
Halothermothrix orenii H 168	2.58	2436	2342	Negative	60
Thermosipho melanesiensis BI429	1.92	1992	1879	Negative	70
Thermotoga lettingae TMO	2.14	2110	2040	Negative	65
Thermotoga petrophila RKU-1	1.82	1864	1785	Negative	80
Thermotoga naphthophila RKU-10	1.81	1867	1768	Negative	80

Rhodothermus marinus DSM 4252	3.39	2965	2863	Negative	65
Saccharomonospora viridis DSM 43017	4.31	3964	3828	Positive	55
Deferribacter desulfuricans SSM1	2.54	2442	2374	Negative	60-65
Thermacetogenium phaeum DSM 12270	2.94	2897	2824	Positive	58

Table 1: List of bacterial species considered to compare the nature of molecular changes between thermophiles and psychrophiles. Bacterial species along with their optimal growth temperature, total number of genes, proteins and the gram stain are presented (http://www.ncbi.nlm.nih.gov/genome/).

2.1.2 Identification of orthologs between the selected thermophiles and psychrophiles:

The BLAST program is used to identify sequence similar homologs from nucleotide or protein databases (Altschul et al. 1990). The program takes a query sequence and searches it against the database selected by the user. It aligns the query sequence against every subject sequence in the database and the results are reported in the form of a ranked list followed by a series of individual sequence alignments, plus various statistics and score parameters. We have performed bi-directional best hit method (BDBH) by offline BlastP program to identify the orthologs for all proteins of 55 thermophiles and 12 psychrophiles. Bi-directional best hit method (BDBH) is widely used method for identification of orthologs in reference organisms, for proteins of a target organism. Bi-directional best-hit method includes identification of gene pairs in two different genomes that are more similar to each other than either is to any other gene (Overbeek et al. 1999). Orthologs for each protein of 55 thermophiles and 12 psychrophiles that are common in all 67 extremophiles were identified. Each protein (query) of one extremophile (an extremophile as query organism) was searched (forward blast) against proteins of rest 66 extremophiles (target organisms) and all best hits of each target organism were reverse blasted (reverse blast) against proteins of query organism. If the top and high scoring hit of reverse blast is the query protein, then the best hits of 66 selected extremophiles of query protein are true orthologs. A blast.pl program was written to automate the bidirectional best hit method, which take each extremophile as query organism one by one and search for the orthologs in rest of the selected extremophiles. Parsing of the blast output files was done by ortho.pl program to check for the presence of true orthologs between the organisms. The program will checks the top and best hits in the forward and reverse blast and if the same homologous proteins are present then they are considered as orthologs. If the homologous proteins are different between the organisms then they are not considered as orthologs. Each protein of an organism with its orthologs in 66 genomes was written into an output file. This output file was again parsed by another thermopsychro.pl program to identify the protein orthologs that are common in all selected thermophiles and psychrophiles. The protein with no orthologs in either of single thermophile or psychrophile was excluded from analysis. Orthologs that are present in all 55 thermophiles and 12 psychrophiles are designated as "conserved proteins". And output file is saved as .txt file. With this method we found 119 proteins are common to all selected thermo and psychrophilic organisms.

2.1.3 Phylogenetic analysis of GroESL proteins:

The phylogenetic tree was constructed using the maximum likelihood method in MEGA 5.0 (Smith et al. 2009). Here, GroES and GroEL protein orthologs in 55 thermophiles and psychrophiles were considered for phylogenetic analysis. The tree was constructed using Jones-Taylor-Thornton (JTT) substitution model. Statistical significance at each node in the trees was evaluated using 1000 bootstrap resamplings.

2.1.4 Amino acid composition analysis:

Amino acid composition analysis was done for 119 conserved proteins identified by bidirectional best hit method.

Generation of input file:

All 119 proteins are concatenated into single protein sequence file using concat.pl program. This single protein sequence file was generated for each 55 thermophiles and 12 psychrophiles which were considered for amino acid composition analysis.

Amino acid composition calculation:

Amino acid composition analysis were done using a *Perl Script* program (aacal.pl) which take each concatenated protein sequence file of an organism and calculates the frequency of occurrence of each residue. Frequency was calculated by counting number of times a particular residue is present divided by the length of the protein sequence and the formula is.

$$\text{Frequency} = \frac{\text{count of individual amino acid} * 100}{\text{total No of amino acid in the protein}}$$

Program is automated which takes concatenated protein files of each organism one by one and writes into an output excel file. Amino acid frequency values of all thermophiles and psychrophiles are clustered into separate files and the mean frequency is calculated. Mean frequency values of thermophiles and psychrophiles were plotted in the graph. To find whether the observed differences in mean frequency values are significant or not between the extremophiles t-test was performed. Unpaired Student t-test with Welch's correction was performed in off line GraphPad Prism 5 software. Significance of the test was done by calculating the *p* value. Similarly amino acid composition analysis and student t-test was performed for the GroESL protein orthologs in thermophiles and psychrophiles.

2.1.5 Calculation of amino acid(s) conserved at a given position of GroESL proteins between thermophiles and psychrophiles:

Multiple sequence alignment of GroES and GroEL protein orthologs in 55 thermophiles and 12 psychrophiles was done by Clustal W (www.ebi.ac.uk/Tools/clustalw/) (Thompson et al. 1994). BLOSUM62 matrix, an open gap penalty of 10 and an extension penalty of 0.05 were the parameters used in multiple sequence alignment. Multiple alignment file was considered for the calculation of amino acid(s) conserved at each position in the alignment between extremophiles. *Perl Script* program was written for which takes *.aln file as input. The main criteria for this analysis includes grouping of amino acids based on their side chain polarity, side chain charge and hydropathy index. Amino acids were grouped as positively charged (Asp (D), Glu (E)), negatively charged (Lys (K), Arg (R)) and neutral hydrophilic (Asn (N), Gln(Q), Gly (G), Pro (P), Ser (S), Thr (T), Trp (W), Tyr (Y)) and neutral hydrophobic (Ala (A), Cys (C), His (H), Ile(I), Leu (L), Met (M), Phe (F), Val (V)). Aligned protein sequence of GroES and GroEL (*.aln file) of thermophiles and psychrophiles was considered as input file. *Perl Script* reads each position in the aligned file and calculates the frequency of occurrence of conserved amino acid at each position between extremophiles according to the formula.

$$\underset{\text{(in the column)}}{\text{Frequency of occurrence of AA}} = \frac{\text{count of individual A.A in the column* 100}}{\text{total No of Proteins}}$$

The amino acids at each position were considered only when there was at least 50 % of occurrence at respective position.

2.1.6 Codon usage bias between the selected extremophiles:

Codon usage bias was done for the open reading frames of 119 proteins with orthologs in all 55 thermophiles and 12 psychrophiles.

Generation of input file:

Codons of each amino acid is grouped based on the third base of triplet codon, if A or T is present then they are grouped as A/T ending codons. Similarly, if G or C is there then they are grouped as G/C ending codons. In this grouping method, methionine and tryptophan were not considered as they have single codon. The count of A/T and G/C ending codons were considered to calculate the frequency of occurrence of A/T or G/C ending codons of particular amino acid. Psychro and thermo orthologs protein files of 119 proteins were considered and *Perl Script* program was written to retrieve gene sequences from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria). For particular organism all 119 gene sequences are concatenated into single gene sequence file using another *Perl Script*. This single gene sequence file was generated for each of the 55 thermophiles and 12 psychrophiles which were considered for codon usage bias.

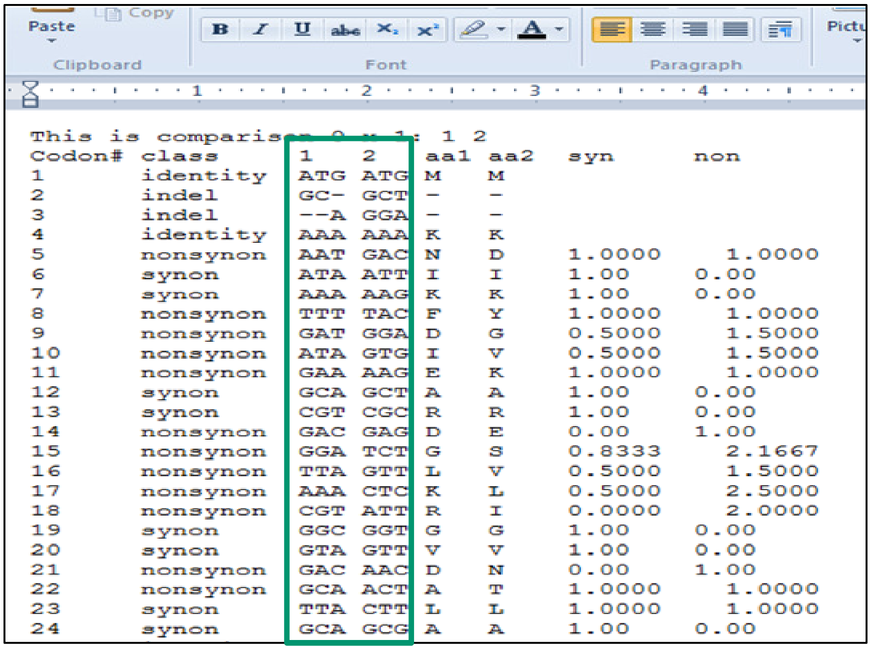
Codon usage calculation:

Codon usage analysis was done by using codoncal.pl program. Concatenated open reading frames of 119 conserved proteins were used as input file. Frequency of occurrence of A/T or G/C ending codons of a particular amino acid is count of A/T or G/C ending codons of particular amino acid divided by the total occurrence of that amino acid. Mean frequency of A/T and G/C ending codons of particular amino acid was calculated by taking the average occurrence of it for a particular selected extremophile. The obtained mean frequency values of thermophiles and psychrophiles were plotted and were used for calculating student t-test to check for the significance of the differences in the mean value between the two groups.

2.1.7 Nucleotide substitution rate analysis:

Synonymous and non-synonymous substitutions were calculated to understand the variation in the nucleotide sequences between the two extremophiles. Synonymous and non-synonymous substitution calculations were done by submitting the aligned nucleotide sequences of 119 conserved proteins of thermophiles and psychrophiles to SNAP (synonymous and non synonymous analysis program). SNAP (http://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html) calculates pairwise synonymous and non-synonymous distances according to the Nei and Gojobori method for an alignment and incorporates statistic calculation developed in Ota and Nei 1994 (Korber 2000; Nei and Gojobori 1986; Ota and Nei 1994). It generates a synonymous and a non-synonymous distance matrix, an exhaustive pairwise codon-by-codon comparison (Fig 2).

Fig 2: Codon by codon aligned file of a sequence pair generated by the SNAP software.



```
This is comparison 0 v 1: 1 2
Codon# class      1   2   aa1 aa2  syn     non
1      identity   ATG ATG M   M
2      indel      GC- GCT -   -
3      indel      --A GGA -   -
4      identity   AAA AAA K   K
5      nonsynon   AAT GAC N   D    1.0000    1.0000
6      synon      ATA ATT I   I    1.00    0.00
7      synon      AAA AAG K   K    1.00    0.00
8      nonsynon   TTT TAC F   Y    1.0000    1.0000
9      nonsynon   GAT GGA D   G    0.5000    1.5000
10     nonsynon   ATA GTG I   V    0.5000    1.5000
11     nonsynon   GAA AAG E   K    1.0000    1.0000
12     synon      GCA GCT A   A    1.00    0.00
13     synon      CGT CGC R   R    1.00    0.00
14     nonsynon   GAC GAG D   E    0.00    1.00
15     nonsynon   GGA TCT G   S    0.8333    2.1667
16     nonsynon   TTA GTT L   V    0.5000    1.5000
17     nonsynon   AAA CTC K   L    0.5000    2.5000
18     nonsynon   CGT ATT R   I    0.0000    2.0000
19     synon      GGC GGT G   G    1.00    0.00
20     synon      GTA GTT V   V    1.00    0.00
21     nonsynon   GAC AAC D   N    0.00    1.00
22     nonsynon   GCA ACT A   T    1.0000    1.0000
23     synon      TTA CTT L   L    1.0000    1.0000
24     synon      GCA GCG A   A    1.00    0.00
```

Fig 2: Codon by codon aligned file of a sequence pair generated by the SNAP software. Codon by codon aligned file between two sequences (1 and 2 here). It provides the information about the synonymous and non-synonymous change (in the second column) for the aligned codons (in third and fourth column). If the aligned codons has a gap or same codon is present then it represent as indels or identity respectively. The amino acids that are encoded by the aligned codons are given in the fifth and sixth column.

Generation of input file:

Psychro and thermo orthologs of 119 proteins were considered and *Perl Script* program was written to retrieve gene sequences from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria). Gene sequences of 119 protein orthologs were aligned using T-coffee software (http://tcoffee.crg.cat/) with BLOSUM62 matrix, an open gap penalty of 10 and an extension penalty of 0.05 were considered. Each gene sequence with its orthologs in 55 thermophiles and 12 psychrophiles were aligned separately and the obtained aligned file was submitted to SNAP which performs pair wise comparison of all sequences. For each pair wise sequence comparison, it generates codon by codon aligned file which gives information about the number of synonymous and non-synonymous substitution (Fig 2). The output codon-codon aligned file was considered as input file for calculation of synonymous (SYN), non-conserved (NCS) (if it results in non-conserved substitution) and conserved non-synonymous (NCNS) (if it results in a conserved substitution) substitution.

Synonymous and non-synonymous substitution calculation:

Perl Script program was written to calculate the nucleotide substitution rate between thermophiles and psychrophiles. It includes three steps (1) calculation of number of synonymous and non-synonymous sites, (2) calculation of number of synonymous and non-synonymous substitution and (3) calculation of substitution rate.

(1) **Calculation of number of synonymous and non-synonymous sites:**

Nusub.pl program was written to read the codon-codon aligned file between two gene sequences generated by the SNAP. The program reads the codon-codons aligned and assign certain values to each position of the triplet codons. In the codon, first and second position is non-synonymous site and third position is mostly synonymous site. Each position in the triplet codon was considered for analysis. If a change in the base in the first, second or third position of a codon with any other base results in change in the amino acid then that site is non-synonymous site and if there is no change in amino acid then that site is synonymous site. Suppose for example, ACG is aligned with TCG, the program check first position in ACG, i.e., A is replaced by any other three bases will result change in amino acid, therefore, it is a non-synonymous site (Table 2). ACG codes for threonine and TCG codes for serine. Threonine and serine are amino acids with similar properties. Therefore, in this example, A to T substitution leads to a conservative amino acid substitution. So the calculation at the first position will be one out of three bases substitution will result in conserved amino acid, then, the program will assign 0.33 (1/3=0.33) and two out of three substitution will result in non conserved amino acid then the program will assign 0.67 (2/3=0.67) (Table 2).

Table 2: Synonymous and non-synonymous sites and the scores assigned to a hypothetical pair wise sequence alignment.

	V			T			R			Q				
Seq1	GTG			ACG			CGC			CAG				
Seq2	GTG			TCG			CGC			CAA				
	V			S			R			Q				
No NCN site	0.33	1	0	0.67	1	0	1	1	0	1	1	0.67	N_{NCNS}	0.33+1+0+.......= 7.67
No CN site	0.67	0	0	0.33	0	0	0	0	0	0	0	0	N_{CNS}	0.67+0+0+......= 1
No SYN site	0	0	1	0	0	1	0	0	1	0	0	0.33	S_{SYN}	0+0+1+........= 3.33

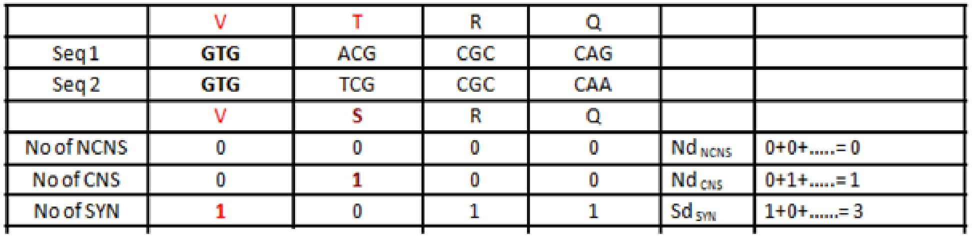
Table 2: Synonymous and non-synonymous sites and the scores assigned to a hypothetical pair wise sequence alignment. Each site of the aligned codons is analyzed and values are assigned by the program based on synonymous and non-synonymous site. In case of non-synonymous site, scores are assigned based on conserved and non-conserved site. Total numbers of synonymous, non-conserved and conserved non-synonymous sites between pair of sequences are tabulated. NCNS, non-conserved non-synonymous; CNS, conserved non-synonymous; SYN, synonymous.

The list of conserved amino acids from the Dayhoff's substitutions was considered for analysis (Dayhoff et al. 1978). Next the program will check second position in ACG and TCG alignment, As substitution of C at this position with any other base results in change of amino acid i.e., non-synonymous site. Then the program will assign value 1 (Table 2). In the third position of codon ACG, substitution of G with any other base will result in no change in amino acid i.e., it is synonymous site, then, the program will assign a value 1 (Table 2). The assigned values for all aligned codons are summed up to get the total number of synonymous sites (S_{SYN}), non-conserved non-synonymous (N_{NCNS}) and conserved non-synonymous site (N_{CNS}) for the particular pair of sequences. All possible pairs of sequences of multiple aligned sequences were analyzed and values are assigned (Table 2).

(2) **Calculation of number of synonymous and non-synonymous substitution:**

In the second step, for the aligned codons, it will check for which amino acid it codes and assign a value 1, if it is synonymous or conserved or non conserved non-synonymous substitution and it assign value 0, if there is same codon and same amino acid (identity) (Table 3). All aligned codons were checked and values assigned based on the type of substitution. The assigned values for all aligned codons were summed up to get the total number of synonymous substitution (Sd_{SYN}), non-conserved non-synonymous (Nd_{NCNS}) and conserved non-synonymous substitution (Nd_{CNS}) for the particular pair of sequences. All possible pairs of sequences of multiple aligned sequences are analyzed and values were assigned (Table 3).

Table 3: Synonymous and non-synonymous substitution and the scores assigned to a hypothetical pair wise sequence alignment.



	V	T	R	Q		
Seq 1	**GTG**	ACG	CGC	CAG		
Seq 2	**GTG**	TCG	CGC	CAA		
	V	**S**	R	Q		
No of NCNS	0	0	0	0	Nd_{NCNS}	0+0+.....= 0
No of CNS	0	**1**	0	0	Nd_{CNS}	0+1+.....= 1
No of SYN	**1**	0	1	1	Sd_{SYN}	1+0+......= 3

Table 3: Synonymous and non-synonymous substitution and the scores assigned to a hypothetical pair wise sequence alignment. Amino acids of aligned codons are compared and scores are assigned based on type of substitutions i.e. synonymous and non-synonymous substitution. In case of non-synonymous substitution, scores are assigned based on conserved and non-conserved substitution. Total number of synonymous and non-conserved and conserved non-synonymous substitutions between the pair of sequences are tabulated. NCNS, non-conserved non-synonymous; CNS, conserved non-synonymous; SYN, synonymous.

(3) **Calculation of substitution rate:**

Fraction of Synonymous (SYN), non-conserved non-synonymous (NCNS) and conserved non-synonymous (CNS) that differ was calculated by

$$d_{SYN} = \frac{\text{No. of synonymous substitutions } (Sd_{SYN})}{\text{No. of synonymous sites } (S_{SYN})}$$

$$d_{NCNS} = \frac{\text{No. of NCNS substitutions } (Nd_{NCNS})}{\text{No. of NCNS sites} (N_{NCNS})}$$

$$d_{CNS} = \frac{\text{No. of CNS substitutions } (Nd_{CNS})}{\text{No. of CNS sites } (N_{CNS})}$$

Substitution rate was calculated by

$$\text{Substitution rate} = \frac{d_{NCNS}}{d_{SYN}}$$

Submission of out put alignment file obtained by SNAP into an automated perl script it generates fraction of synonymous, non-conserved nonsynonymous and conserved non-synonymous sites and substitutions. Average was calculated for the d_{SYN}, d_{NCNS} and d_{CNS} for that gene sequence. Similarly, all 119 gene sequences with its orthologs were analyzed and the obtained average values of d_{SYN}, d_{NCNS} and d_{CNS} were plotted.

For the GroESL protein, gene sequences of GroES and GroEL were aligned separately with its orthologs and d_{SYN}, d_{NCNS} and d_{CNS} for each pair sequences were calculated. The obtained values of d_{SYN}, d_{NCNS} and d_{CNS} were plotted and compared between extremophiles.

2.1.8 Retrieval of non-coding upstream sequences of *groES*:

We retrieved a maximum of 500 bp upstream of *groES* or the actual intergenic DNA regions, if the distance between the two transcriptional units is less than 500 bases. Intergenic DNA sequences were retrieved from NCBI using a *Perl Script*. The out put file was saved into another .txt file for the identification of conserved inverted repeats, direct repeats and small oligonucleotide motifs.

2.1.9 Identification of *cis*-regulatory elements:

The upstream DNA sequence of *groES* was taken as input to Motif extraction by multiple expectation maximization (MEME ver. 4.3.0) which is an EM based algorithm (Bailey and Elkan 1994). Intergenic sequences of orthologs were given as input to MEME to detect *cis* regulatory elements. In case of intergenic region of less than 30 bp were not considered for analysis. The default parameters used for the identification of the motif by MEME include a repeat of minimum length of 6 bp and maximum of 30 bp nucleotide and loop region of less than 10 nucleotide length.

2.1.10 Construction of consensus sequence of GroESL:

Consensus sequences of GroES and GroEL for a group of 12 psychrophiles and 55 thermophiles were constructed by aligning protein sequences and by calculating the frequency of occurrence of a residue at each position. The amino acid residue at each position in a consensus sequence was considered only when there was at least 50% of occurrence at respective position.

2.2 *In vitro* analysis

2.2.1 Synthetic-*groESL* gene synthesis:

These consensus GroESL proteins were reverse translated using the *E. coli* genetic code and the synthetic psychro-*groESL* and thermo-*groESL* were generated. Consensus sequences of thermo and psychro GroES and GroEL were given to GenScript USA Inc. (Piscataway, New Jersey, USA) for synthesis (Fig 3).

Fig 3: Schematic representation of the synthetic genes construct of thermo and psychro GroES and GroEL cloned into pUC57.

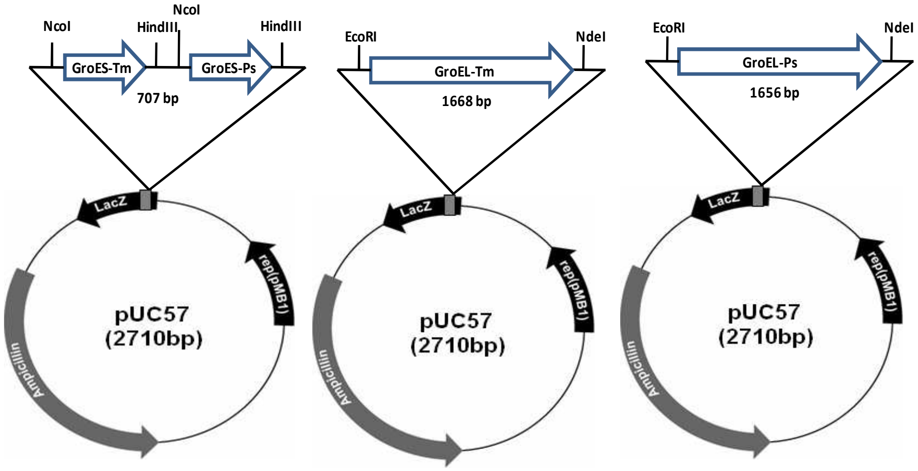


Fig 3: Schematic representation of the synthetic genes construct of thermo and psychro GroES and GroEL cloned into pUC57. Thermo and psychro GroES and GroEL synthetic genes were cloned into pUC57 vector by Gene Script Inc. Tm, thermophiles; Ps, psychrophiles. Restriction enzyme sites shown are used for cloning the open reading frames in to an expression vector pET28a(+) vector for expression of the protein.

2.2.2 Primer design for expression of synthetic *groESL* genes:

The synthesized *groESL* genes of thermophile and psychrophile were cloned into pUC57 vector separately by GeneScript with the mentioned restricted enzyme sites so that synthetic genes can be excised from the pUC57 vector (Fig 3). Consensus g*roES* genes of thermophile and psychrophile were merged together with the required restriction enzyme sites and incorporated within the sequence before given for synthesis at the C and N terminal site of the ORF (Fig 3). The required *groES* ORF can be amplified by the designed expression primers with *NcoI* and *HindIII* restriction site in forward and reverse primer respectively (underlined in primer sequence). *NcoI* and *HindIII* restriction sites were selected for insertion of *groES* gene into pET28a(+) vector for the generation of C-terminal HIS tagged GroES protein. Similarly, the expression primers are designed for the GroEL of thermophile and psychrophiles with *EcoRI* and *NdeI* restriction site in forward and reverse primer respectively before giving for synthesis (underlined in primer sequence). *EcoRI* and *NdeI* restriction sites were considered for generation of N-terminal His-tag GroEL protein.

Thermo-GS-F-EXP:
3’ GTA<u>CCATGG</u>AAATTCGCCCGCTGGGCGATC 5’
Thermo-GS-R-EXP:
3’ GGTC<u>AAGCTT</u>TTCTTCAATCACCGCCAGAATATCG 5’
Psychro-GS-F-EXP:

3’ GTACCATGGACATTCGCCCGCTGCATGAT 5’

Psychro-GS-R-EXP:

3’ GGTCAAGCTTTTCCACAATCGCCAGAATATCGCT 5’

Psychro_GL_F_Exp:

3’ GACTCATATGGCGGCGAAAGAAGTGCTG 5’

Psychro_GL_R_Exp:

3’ CGCTGAATTCACATCATGCCGCCCATGCC 5’

Thermo_GL_F_Exp:

3’ CACTCATATGGCGAAACAGATTATTTTTGATG 5’

Thermo_GL_R_Exp:

3’ TGCTGAATTCAAAAATCCATATCGCCCGCG 5’

2.2.3 Cloning of synthetic *groESL* genes into pET28a(+) vector:

pET28a(+) expression vector with kanamycin resistance marker was considered for our expression studies of synthetic *groESL* genes in *Roseta* strain of *E. coli*. Synthetic *groEL* genes of the thermophiles and psychrophiles were digested with *NdeI* and *EcoRI* enzymes to extract from pUC57 vector in presence of buffer O at 37 ^{o}C for 16 h. pET28a(+) vector was also double digested with *NdeI* and *EcoRI* enzymes. The digested samples of synthetic *groEL* genes were eluted from the gel by gel extraction kit method. The isolated piece of gel was dissolved in an adjusted volume of solution buffer at 55 °C. The contained DNA was then bound to a silicon column, washed, dried and eluted in 25 to 50 µL of elution buffer. The digested and purified sample of *groEL* genes of thermophile and psychrophile were ligated with the digested pET28a(+) vector in 4:1 ratio at 16 °C with overnight incubation.

Synthetic *groES* genes of thermophile and psychrophile were amplified from the recombinant pUC57 vector with *Pfu* DNA polymerase enzyme with 58 °C annealing and 70 °C final extension temperatures. The amplified fragments of *groES* genes were purified by phenol-chloroform-isoamyl alcohol method and concentration was checked. The purified samples of *groES* genes of thermopliles and psychrophiles were digested with *NcoI* and *HindIII* restriction enzymes. Ligation was setup with the digested and purified samples of GroES of thermophiles and psychrophiles with pET28a(+) vector with 4:1 ratio at 16 °C with overnight incubation. The ligated mixture was transformed into *E. coli* DH5α competent cells and plated on LB agar containing kanamycin (25 ug/ml) and positive colonies are screened.

2.2.4 Expression and purification of synthetic GroESL protein:

The N-terminally His-tagged psychro-GroEL and thermo-GroEL proteins were expressed separately in BL21(DE3)pLysS, which had been transformed with pET-psy-GroEL and pET-ther-GroEL and was purified using HIS-Select^{TM} Nickel Affinity gel (SIGMA - P6611) according to the supplier's instructions. The expression of protein was induced by addition of 400 µM (Final concentration) IPTG. Bacterial cells were collected by centrifugation at 10,000 *g* for 10 min and pelleted cells were disrupted with a sonic oscillator (Model, UV2070, probe MS-72, Bandelin Electronic, Berlin) operated for 10 min at 50 % power, with 1 min pulse interval, in 100 mM Tris-HCl (pH 8.0) and 200 mM NaCl. Insoluble materials were removed by centrifugation at 20,000 *g* for 20 min at 4^{o}C. The supernatant was loaded onto a HIS-Select ^{TM} Nickel Affinity column. The column was washed with 50 mM NaH_2PO_4 (pH 7.5), 200 mM NaCl and 10 mM imidazole and sequentially with the same buffer containing 40 mM imidazole. Then His tagged psychro-GroEL and thermo-GroEL was eluted with 50 mM NaH_2PO_4 (pH 7.5), 200 mM NaCl and 200 mM imidazole. The purity of each fraction was examined by SDS gel electrophoresis. The fractions which gave a single protein band at the expected region on the gel were combined and dialysed against 50 mM NaH_2PO_4 (pH 7.5) using amicon ultra filters (Millipore-UFC800324). The resultant protein was used for thermal stability analysis.

2.2.5 Thermal denaturation studies of GroESL protein:

Thermophilic and Psychrophilic GroEL proteins, made out of conserved sequences were tested for their thermal stability. Psychro and thermo GroEL in 50 mM Tris buffer, at pH 8 with the approximate concentration of 0.15 µM and 2.4 µM, respectively were used for the thermal denaturation studies. Temperature of the proteins was increased from 10 °C to 90 °C using a Peltier with the rate of 1 °C/min. The secondary structural changes were followed using Jasco J-810 spectropolarimeter at 220 nm wavelength with the path length of 2 mm and bandwidth of 1 nm. The obtained protein ellipticity values were normalized and fitted using Gibbs-Helmholtz equation (Kumar et al. 2006) with an assumption that the proteins follow a cooperative two-state unfolding process.

Chapter 3:

3.1 Introduction:

Comparative studies were performed to understand how the protein and DNA molecules attain stability for the organisms that are adapted to extreme conditions (Gromiha et al. 1999; Metpally and Reddy 2009; Singer and Hickey 2003; Basak et al. 2004; Basak and Ghosh 2006; Zhou et al. 2008). Reports are available on strategies adapted by microorganisms to survive extreme temperature conditions. Metpally and Reddy, (2009), compared amino acid composition between psychrophilic and mesophilic proteins to understand the contribution of amino acids in proteins to low temperature adaptation (Metpally and Reddy 2009). Amino acids Gly, Ser, Lys and Asp in mesophilic proteins were substituted by Ala, Thr, Arg and Glu in thermophilic proteins to enhance thermal stability of proteins (Vogt and Argos 1997). The substitution of one amino acid with other has a greater chance of being acceptable based on the substitution frequencies between the amino acids (Dayhoff et al. 1972). Conserved amino acid substitutions are the favored changes in protein sequences based on their physiological properties, molecular bulk side chain, hydrophobicity etc., while non-conserved substitutions are un-favored changes in the protein sequence. Basak et al. (2004), compared 529 orthologous protein sequences between a mesophile (*B. subtilus*) and a thermophile (*A. aeolicus*) (Basak et al. 2004). They showed that amino acids Phe, Leu, Ile, Val, Pro, Tyr, Lys, Glu, Trp and Arg occurred with relatively higher frequency in thermophiles when compared to mesophiles (Basak et al. 2004). In another study, McDonald et al. (1999), observed that in a thermophile (*Methanococcus*) occurrence of Glu, Ile, Lys, Met, Asn, Gln, Arg, Ser and Thr were relatively high when compared to a mesophile of same genus (McDonald et al. 1999). Goldstein (2007), observed that there were higher levels of hydrophobic residues (Ala, Cys, Gly, Val, and Met) and lower levels of charged and aromatic residues in psychrophilic proteins when compared to mesophilic proteins by quasi chemical approximation (Goldstein 2007). During the course of adaptation to extreme temperatures, depending upon the temperature the microorganisms has been constantly exposed to, point accepted mutations in the genetic material which might have led to the selected changes in amino acids for optimum enzymatic function.

According to Basak et al. (2004) and McDonald et al. (1999), residues Ser, Thr, Gln and Asn are found to occur more in thermophiles compared to mesophiles, but residues Gln and Asn undergo oxidation at high temperature (Basak et al. 2004; McDonald et al. 1999; Russell et al. 1994). Ser, Thr and other polar residues are water interacting residues that form the advantageous parameter for psychrophilicity (Jahandideh et al. 2008). Threonine is more in anti freeze proteins and is essential for preventing ice nucleation (Chou 1992). Studies between thermophiles and psychrophiles are essential to understand the differences which might be accurate and specific to that particular extremophilic group. No reports are available till today and it is of utmost importance to understand the molecular adaptation of these two extremophilic groups. In our analysis, all thermophiles and psychrophiles for which the genome sequence information is available till 2009 were considered in our analysis. The rational of the study is to identify those important residues that contribute to the stability of a protein molecule and gain insight into the molecular evolution of thermophiles and psychrophiles to either extreme temperature.

Two distant classes of extremophiles i.e., thermophiles and psychrophiles were considered in this present study. Mesophiles were not considered for analysis, as available information on mesophiles is so vast that on including mesophiles, can make the analyses cumbersome. Comparative analysis of conserved orthologous proteins between thermophilic and psychrophilic bacteria is necessary. As it may reveal differences at the molecular level and these differences might be because of their adaptation to extreme conditions. Molecular level changes are the gradual changes in the protein sequences that occur thorough the natural evolutionary process and by point mutations (Romero and Arnold 2009).

3.2 Results and Discussion:

We have considered 12 psychrophiles and 55 thermophiles in our study for which the genome sequence information was available from NCBI (ftp//:ncbi.nlm.nih.gov/genbank/genomes). List of bacterial species considered for comparative analysis and the temperature ranges at which they had been adapted to survive are given in Table 1. Table 1 gives the information about total number of genes, and proteins of extremophiles used for the current study. Only one strain was considered, when extremophiles with same species and genus names, but with difference in strain names. Most of the genomes of the extremophiles are composed of singular chromosome but *Pseudoalteromonas haloplanktis, Aliivibrio salmonicida* and *Photobacterium profundum* has two circular chromosomes (Médigue et al. 2005; Hjerde et al. 2008; Vezzi et al. 2005). Genome size varies between thermophiles and psychrophiles. An average genome size of thermophiles is 2.68 Mb and psychrophiles it is 4.23 Mb. Average genome size of thermophiles is smaller compared to psychrophiles which might be due to natural selection and mutational rate (Lee and Marx 2012). During adaptation to their respective habitat, organisms undergo reduction in the genome size by deletion of paralogous genes (gene loss) or unnecessary genes (Kettler et al. 2007). *Prochlorococcus* which grows at high and low light niches has smaller genome compared to *Synechococcus* that grows in marine water, which might be due to the gene loss in *Prochlorococcus* rather than gene gain in *Synechococcus*. Smaller genome size was due to the loss of entire gene families and deletion of paralogous genes in *prochlorococcus* (Kettler et al. 2007).

3.2.1 To identify the proteins those are common (conserved orthologs) in all thermophilic and psychrophilic bacterial genomes.

All 55 thermophiles and 12 psychrophiles were used for finding proteins that are common to all 67 extremophiles selected using bidirectional best hit method as described in methodology. Each organism was found to show variation in number of conserved proteins, for example *P. haloplanktis* (as query organism), showed a total of 119 conserved protein hits, where as *T. fusca* (as query organism) showed a total of 239 conserved protein hits and so on. This might be because of the presence of iso forms of proteins. We considered the organism with the lowest number of conserved protein hits, i.e., *P. haloplanktis* as this will not include the isoforms of proteins. 119 conserved proteins of *P. haloplanktis* were listed out in the Table 4. All 119 proteins were clustered using thermopsychrocluster.pl program into two files one containing psychro-orthologs and other with thermo-orthologs. We used these 119 conserved proteins and corresponding open reading frames for data generation in this study. To identify the differences in occurrence of amino acids at various positions of a conserved protein between extremophiles, GroES and GroEL proteins (here after GroESLs) were selected out of 119 conserved proteins.

Table 4: List of 119 proteins from *P. haloplanktis* TAC125, which are commonly present in all thermophiles and psychrophiles selected for this study.

1	77358983	DNA replication initiator	61	77360342	50S ribosomal protein L20
2	77358986	DNA gyrase subunit B	62	77360364	DNA gyrase subunit A
3	77359004	10-formyltetrahydrofolate:L-methionyl-tRNA(fMet) N-formyltransferase	63	77360652	seryl-tRNA synthetase
4	77359005	peptide deformylase	64	77360663	translation initiation factor IF-1
5	77359115	30S ribosomal protein S10	65	77360748	3-oxoacyl-ACP reductase
6	77359116	50S ribosomal protein L3	66	77360870	aspartate tRNA synthetase
7	77359117	50S ribosomal protein L4	67	77360939	DNA polymerase III subunit alpha
8	77359119	50S ribosomal protein L2	68	77360940	RNAse HII
9	77359120	30S ribosomal protein S19	69	77360953	membrane-associated protease
10	77359121	50S ribosomal protein L22	70	77360956	UDP pyrophosphate synthetase
11	77359122	30S ribosomal protein S3	71	77360958	uridylate kinase
12	77359123	50S ribosomal protein L16	72	77360960	30S ribosomal protein S2
13	77359125	30S ribosomal protein S17	73	77360961	methionine aminopeptidase
14	77359138	glutamine synthetase	74	77360988	bifunctional: 5,10-methylene-tetrahydrofolate

					dehydrogenase; 5,10-methylene-tetrahydrofolate cyclohydrolase
15	77359188	transcription antitermination factor	75	77360989	cysteinyl-tRNA synthetase
16	77359189	50S ribosomal protein L11	76	77360997	bifunctional: folylpolyglutamate synthase; dihydrofolate synthase
17	77359190	50S ribosomal protein L1	77	77361205	phosphopantetheine adenylyltransferase
18	77359192	50S ribosomal subunit protein L7/L12	78	77361227	Gcp glycoprotease
19	77359193	DNA-directed RNA polymerase subunit beta	79	77361292	serine hydroxymethyltransferase
20	77359194	DNA-directed RNA polymerase subunit beta'	80	77361341	valine tRNA synthetase
21	77359196	30S ribosomal protein S7	81	77361355	50S ribosomal protein L9
22	77359197	elongation factor G	82	77361356	30S ribosomal protein S18
23	77359226	molecular chaperone GroEL	83	77361414	cell division protein ftsZ
24	77359227	Chaperonin	84	77361417	D-alanine--D-alanine ligase
25	77359316	DNA primase	85	77361418	UDP-N-acetyl-muramate:alanine ligase, L-alanine adding enzyme
26	77359320	cell division transporter substrate-binding protein	86	77361421	UDP-N-acetylmuramoylalanine--D-glutamate ligase
27	77359434	UDP-galactose 4-epimerase	87	77361422	phospho-N-acetylmuramoyl-pentapeptide transferase
28	77359489	alanyl-tRNA synthetase	88	77361427	S-adenosyl-methyltransferase MraW
29	77359549	phosphoglycerate kinase	89	77361436	Methyltransferase
30	77359588	glutamyl-tRNA synthetase	90	77361451	periplasmic serine endoprotease
31	77359609	arginyl-tRNA synthetase	91	77361453	UDP-N-acetylglucosamine 1-

					carboxyvinyltransferase
32	77359646	RecA protein (recombinase A)	92	77361523	inosine/xanthosine triphosphatase
33	77359681	GTP-binding protein LepA	93	77361524	oxygen-independent coproporphyrinogen III oxidase
34	77359692	CTP synthetase	94	77361550	dimethyladenosine transferase
35	77359693	phosphopyruvate hydratase	95	77361569	GTPase ObgE
36	77359790	SsrA-binding protein	96	77361570	50S ribosomal protein L27
37	77359819	cell division protease	97	77361571	50S ribosomal subunit protein L21
38	77359862	30S ribosomal protein S20	98	77361620	single-strand binding protein (SSB) (helix-destabilizing protein)
39	77359865	isoleucyl-tRNA synthetase	99	77361622	DNA excision repair enzyme subunit, with UvrBC
40	77359880	23S rRNA pseudouridine synthase D	100	77361625	D-ribulose-5-phosphate 3-epimerase
41	77359889	GTP-binding protein	101	77361658	preprotein translocase subunit SecA
42	77359890	30S ribosomal protein S16	102	77361672	multifunctional DNA polymerase I: 5'->3' exonuclease/3'->5' polymerase/3'->5' exonuclease
43	77359893	50S ribosomal protein L19	103	77361700	guanylate kinase
44	77359942	transcription elongation factor NusA	104	77361710	DNA-directed RNA polymerase subunit alpha
45	77359945	tRNA pseudouridine 5S synthase	105	77361711	30S ribosomal protein S4
46	77359946	30S ribosomal protein S15	106	77361714	preprotein translocase subunit SecY
47	77359947	polynucleotide phosphorylase/polyadenylase	107	77361715	50S ribosomal protein L15
48	77359977	leucyl-tRNA synthetase	108	77361717	30S ribosomal protein S5
49	77359992	involved in methylthiolation of isopentenylated A37 derivatives in tRNA	109	77361718	50S ribosomal protein L18
50	77359994	GTP-dependent nucleic acid-binding	110	77361720	30S ribosomal protein S8

		protein EngD			
51	77359995	peptidyl-tRNA hydrolase	111	77361722	50S ribosomal protein L5
52	77359997	ribose-phosphate pyrophosphokinase	112	77361724	50S ribosomal protein L14
53	77360001	peptide chain release factor 1	113	77361895	L-glutamine:D-fructose-6-phosphate aminotransferase
54	77360053	excinuclease ABC subunit B	114	77361918	preprotein translocase subunit YidC
55	77360144	DNA polymerase III subunits gamma/tau	115	77362056	histidyl-tRNA synthetase
56	77360151	adenylate kinase	116	77362268	molecular chaperone DnaK
57	77360165	nucleotide exchange factor	117	162139923	excinuclease ABC subunit C
58	77360260	methionyl-tRNA synthetase	118	162139924	phenylalanyl-tRNA synthetase subunit alpha
59	77360266	aminotransferase AlaT	119	162139925	phenylalanyl-tRNA synthetase subunit beta
60	77360339	threonyl-tRNA synthetase			

Table 4: List of 119 proteins from *P. haloplanktis* TAC125, which are commonly present in all thermophiles and psychrophiles selected for this study. 119 conserved proteins of *Pseudoalteromonas haloplanktis* TAC125 are listed with their Gene I.D and the name of the protein. These were generated by BDBH method. These protein sequences were used for calculating amino acid composition. Corresponding open reading frames of these proteins were used for codon bias and synonymous and non-synonymous substitution analysis.

3.2.2 Phylogenetic analysis of GroESL of thermophiles and psychrophiles:

Phylogenetic analysis of the GroES and GroEL proteins showed two clusters, one representing all psychrophiles and another representing most of the thermophiles (Fig 4 and 5). There may be specific differences in the amino acids between these two clusters and such specific differences are probably conserved within each cluster and might be responsible for separation of all orthologs of a protein into two major clusters. GroELs are separated into two different clusters, in spite of having highly conserved amino acid regions throughout the length of protein sequence, irrespective of their temperature adaptation. Though the GroESL proteins belong to protein folding machinery and perform the same function (i.e., folding of nascent or denatured proteins) in either extremophiles, their optimum temperature of refolding activity is different (Ferrer et al. 2003). These specific differences between thermophilic and psychrophilic GroESLs are conserved within each cluster, supporting the argument that during the course of adaptation to extreme conditions, selective changes in genetic material might have led to changes in amino acids for optimized protein function. Such specific differences in amino acids might have arisen during bacterial adaptation to extreme temperatures. The specific amino acid differences, nature of amino acids and their interactions are critical for GroESL temperature optima as evidenced from the comparative analysis performed in this study.

Fig 4: Phylogenetic tree of GroES of thermophiles and psychrophiles.

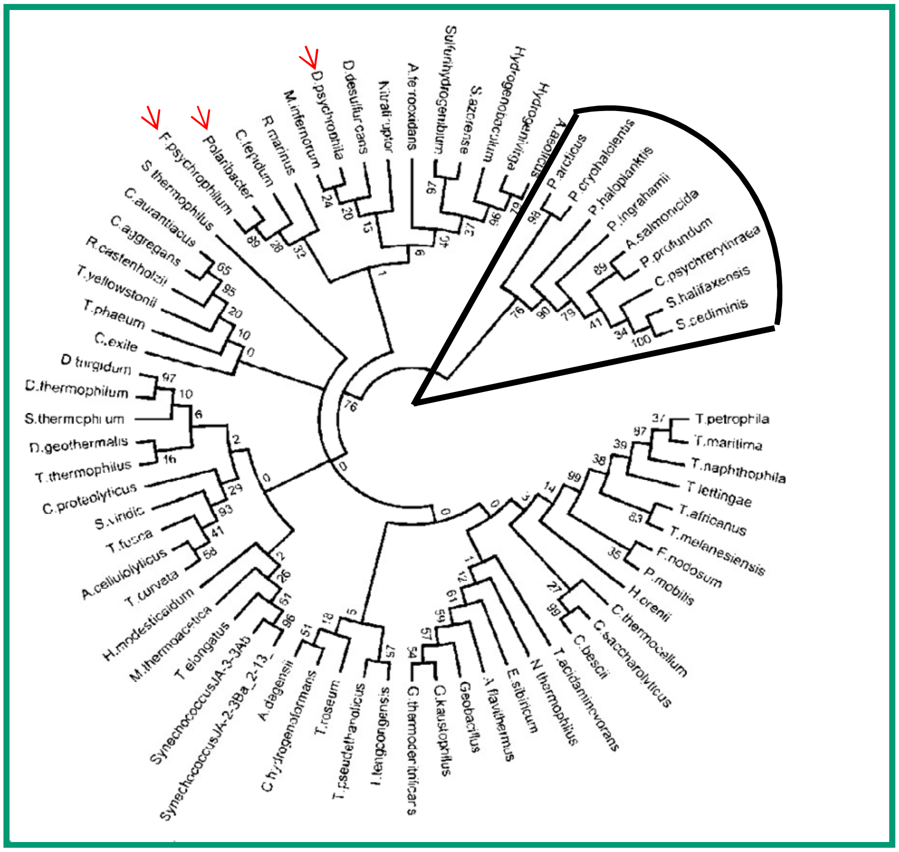


Fig 4: Phylogenetic tree of GroES of thermophiles and psychrophiles. The phylogenetic tree was constructed by Maximum Likelihood Method using MEGA 5.0 software. Numbers on the branches represent the percentage of 1000 bootstrap samples supporting the branch.

Fig 5: Phylogenetic tree of GroEL of thermophiles and psychrophiles.

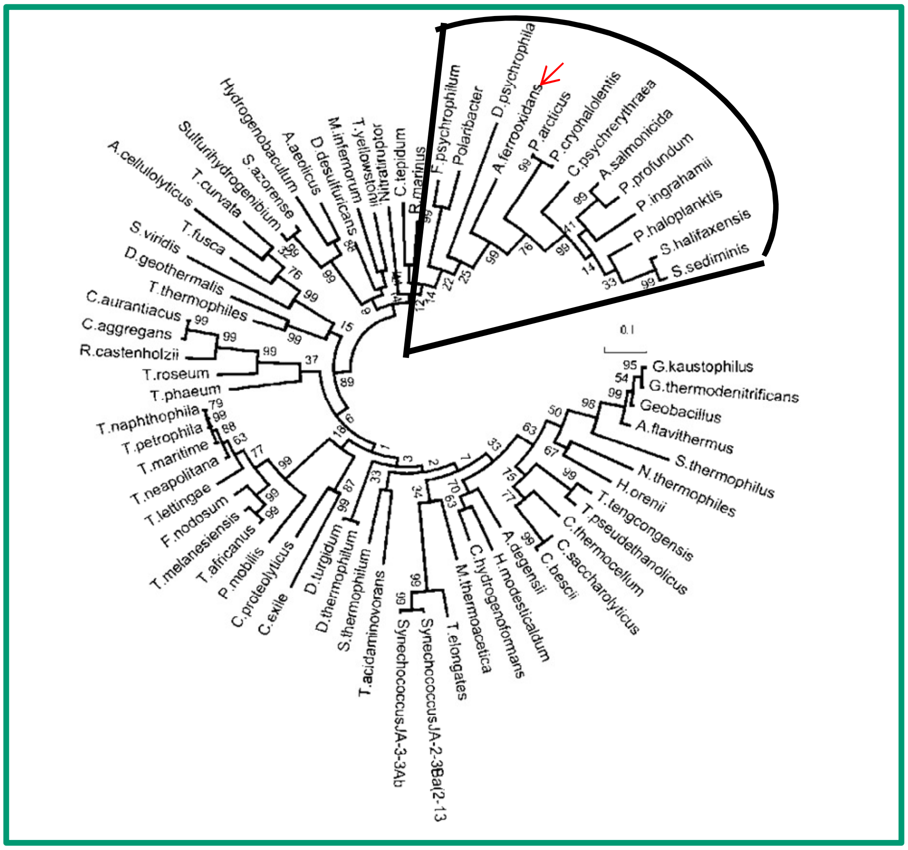


Fig 5: Phylogenetic tree of GroEL of thermophiles and psychrophiles. The phylogenetic tree was constructed by maximum likelihood method using MEGA 5.0 software. Numbers on the branches represent the percentage of 1000 bootstrap samples supporting the branch.

3.2.3 To identify the differences in the amino acid composition between thermophilic and psychrophilic conserved orthologs that might have contributed to the stability of proteins at high or low temperature:

Proteins from the thermophiles are stable, rigid and more active at high temperature than the homologous protein in mesophiles, which grow at moderate temperature (Kumar 2000). Much research was done to identify the reasons for enhanced thermostability by performing the comparative studies between thermophiles and mesophiles (Pack and Yoo 2004; Gromiha et al. 1999; Gromiha and Suresh 2008; Fujita and Kanehisa 2005). No reports are available between thermophiles and psychrophiles, as these two groups belong to two extreme temperature conditions. The analysis on these two extremophiles might provide accurate results, and the observed differences in the amino acid composition might be directly or indirectly involved in protein stability at their respective temperature. Amino acid composition analysis of 119 conserved orthologs showed significant differences in the composition of amino acids between thermophiles and psychrophiles. Student t-test was performed to check for the significant differences in amino acid composition between thermophiles and psychrophiles. Significant student t-test values thus obtained are shown in bold font in Table 5. The obtained results are very much coinciding with the published literature based on the comparative study between thermophiles and mesophiles or among thermophiles, mesophiles and psychrophiles.

Amino acid composition analysis showed that aliphatic amino acids are equally distributed between thermophilic and psychrophilic proteins (Fig 6, Table 5). Aliphatic amino acids are small residues with short alkyl group that are involved in hydrophobic interactions and are the main force for maintaining conformational stability in the inner part of the protein (Creighton 1997). These amino acids with short alkyl group interact more closely with neighboring residues thus resulting in better packed protein structure (Zhou et al. 2008). Among the negatively charged amino acids aspartic acid were seen more in psychrophilic proteins while glutamic acid was more in thermophilic proteins (Fig 6, Table 5). Aspartic acid is unstable at high temperature and its occurrence was low in hyper-thermophiles (Szilagyi and Zavodszky 2000). But results of Tanaka *et al.* 2004 showed that interactions of aspartic acid with other residues are critical for protein stability in hyper-thermophiles which contradicts our finding (Tanaka et al. 2004). Positively charged amino acids lysine, arginine except histidine was seen more in thermophilic proteins than psychrophilic proteins (Table 5). Presence of more charged amino acids in thermophiles than psychrophiles infers that charged amino acids are necessary for entropically more stable conformation of proteins (Osváth et al. 2007). Stable folded conformation was possible by non-covalent interactions like electrostatic, ionic, vander-waal and hydrogen bonds which are mainly due to charged atoms on the protein molecule. High occurrence of charged residues in thermophiles infers that charged residues participates in interactions with other residues resulting in thermal stability of proteins at high temperature. Whereas less frequency of charged residues in psychrophiles infers reduction in number of interactions, thereby, resulting in structure flexibility (Feller 2013). More flexible structure favors more number of substrate molecules to be converted to products by increasing the catalytic activity (Feller 2013). More charged residues were reported in hyper-thermophilic proteins than mesophilic proteins except histidine, which has lower frequency in thermophilic proteins (Chakravarty and varadarajan 2000; Pack and Yoo 2004). Using machine learning algorithm, it was shown that charged amino acids and hydrophobic residues valine and isoluecine have higher occurrence in thermophilic proteins than mesophilic proteins (Gromiha and Suresh 2008).

Occurrence of aromatic amino acids (Phe, Trp and Try) was higher in thermophilic proteins than psychrophilic proteins (Fig 6). It was well known that aromatic amino acids are involved in cation-pi interactions that are necessary for maintaining conformational stability of protein molecule (Ma and Dougherty 1997).

Fig 6: Mean frequency of occurrence of each amino acid between thermophilic and psychrophilic conserved proteins.

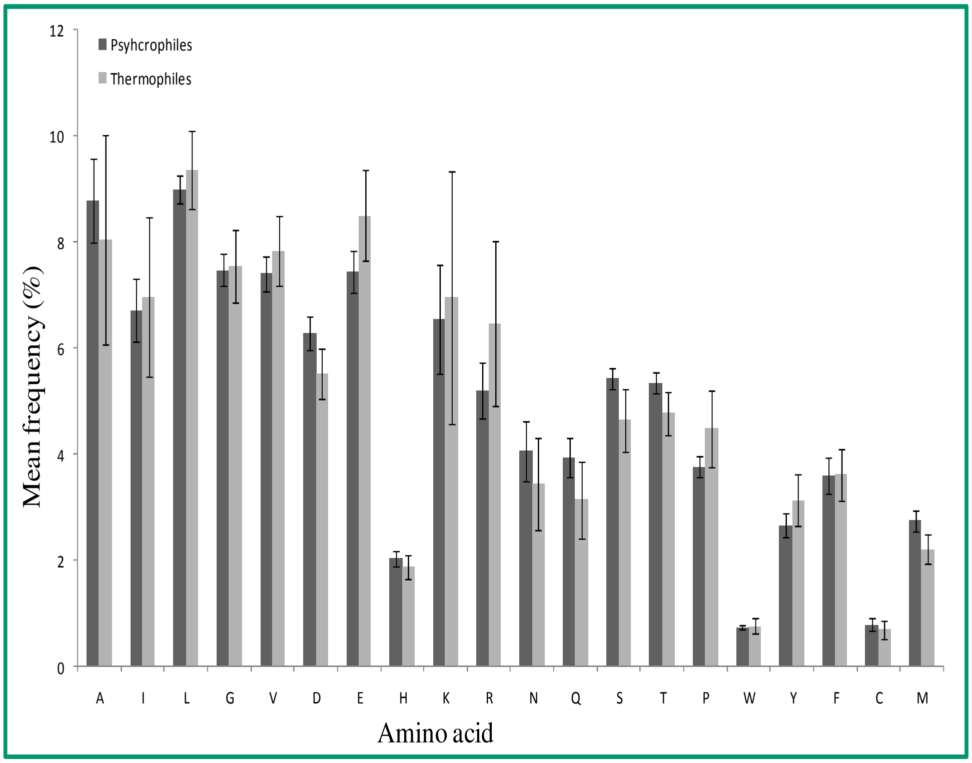


Fig 6: Mean frequency of occurrence of each amino acid between thermophilic (grey bar) and psychrophilic (black bar) conserved proteins. It shows significant differences in the composition of amino acids between thermophiles and psychrophiles. Amino acids are represented in single letter codes on the *X*-axis and the mean frequency (in percentage) of each amino acid occurrence is represented on the *Y*-axis.

Comparative analysis of 24 protein families for which the crystal structure information was available showed that thermophilic proteins have a large number of pair-wise aromatic interactions than its mesophilic homologue (Kannan and Vishveshwara 2000). Proline content is more in thermophilic proteins than psychrophilic proteins (Fig 6). Proline is known to stabilize the folded conformation in most thermophilic proteins by decreasing the entropy of unfolding, thereby stabilizing the secondary structural elements (Matthews et al. 1987). Due to its pyrolidine ring, it adopts few configurations with lowest conformational entropy which results in rigid conformation (Watanabe et al. 1997). Many reports showed that thermophilic proteins have higher frequency of proline and it was used to increase the protein stability by several mutational studies (Xu et al. 2003; Pack and Yoo 2004; Sadeghi et al. 2006; Veltman et al. 1996). The observed differences in amino acid composition were very much co-rrelating with the already published work on the comparative studies between thermophilic and mesophilic proteins (Metpally and Reddy 2009; Pack and Yoo 2004; Gromiha et al. 1999; Zhang and Fang 2006; Fukuchi and Nishikawa 2001; Fujita and Kanehisa 2005; Goldstein 2007).

Table 5: Normalized frequency of each amino acid in the conserved proteins between thermophiles and psychrophiles.

	Psychrophiles	Thermophiles	
Amino acids	Mean±SD	Mean±SD	t-test (*p*-value)
Ala	8.780±0.784	8.044±1.982	2.100 (0.0414)
Ile	6.713±0.584	6.968±1.492	-0.971 (0.3366)
Lue	9.001±0.268	9.372±0.738	-2.949 (0.0049)
Gly	7.471±0.303	7.547±0.696	-0.595 (0.5549)
Val	7.409±0.326	7.837±0.649	-3.328 (0.0022)
Asp	6.283±0.306	5.517±0.464	**7.083 (< 0.0001)**
Glu	7.442±0.399	8.504±0.850	-**6.530 (< 0.0001)**
His	2.043±0.144	1.883±0.221	3.131 (0.0045)
Lys	6.551±1.029	6.954±2.385	-0.921 (0.3625)
Arg	5.213±0.526	6.477±1.555	-**4.882 (< 0.0001)**
Asn	4.066±0.569	3.445±0.873	3.069 (0.0054)
Gln	3.942±0.371	3.138±0.735	**5.510 (< 0.0001)**
Ser	5.433±0.204	4.638±0.588	**8.056 (< 0.0001)**
Thr	5.347±0.197	4.771±0.418	**7.205 (< 0.0001)**
Pro	3.768±0.196	4.481±0.719	**-6.360 (< 0.0001)**
Trp	0.734±0.038	0.767±0.138	-1.544 (0.1277)
Tyr	2.659±0.226	3.129±0.484	-**5.085 (< 0.0001)**
Phe	3.602±0.343	3.618±0.491	-0.133 (0.8955)
Cys	0.790±0.124	0.700±0.173	2.114 (0.0466)
Met	2.752±0.200	2.209±0.281	**7.877 (< 0.0001)**

Table 5: Normalized frequency of each amino acid in the conserved proteins between thermophiles and psychrophiles. Frequency of each amino acid along with their standard deviation (SD) were mentioned. Student t-test shows the significant differences in amino acid composition between thermophiles and psychrophiles. Significant compositional differences are highlighted in bold between extremophiles.

Uncharged polar residues (Asn, Gln, Ser, Thr, Cys and Met) were seen more in psychrophilic proteins than thermophilic proteins (Fig 6). Residues Gln, Asn, Cys and Met are thermoliable, Gln and Asn undergo deamination while Cys and Met under go oxidation at high temperatures (Catanzano et al. 1997; Russell et al. 1994). It was reported that thermophilic proteins have lesser number of methionine residues than mesophilic proteins (Kumar 2000; Xu et al. 2003). Beta glucosidase A from thermophilic bacteria was found to have lower frequency of methionine than mesophilic bacteria (Lopez-Camacho et al. 1996). Gromiha et al. (1999), performed comparative analysis between mesophiles to thermophiles and found that the strong preference of replacements was seen from Met-Ala, Cys-Ala, Trp–Try, Met–Leu, Cys-Val and Cys-Ile from mesophiles to thermophiles (Gromiha et al. 1999). Five out of six replacements were from Cys and Met residue in mesophlies to other hydrophobic residues in thermophiles. Low occurrence of Cys residue was observed in thermophilic proteins in their comparative study with mesophilic proteins (Kumar et al. 2000; Xu et al. 2003). Ser and Thr interact with water present surrounding the protein molecule (Mattos 2002). At high temperatures the interactions break as water molecule that interacts with Ser and Thr would be released, thereby, destabilizing the water binding site resulting in protein instability (Denisov et al. 1997;

Nagendra et al. 1998). Ser or Thr or any other polar residue whose side chain can form a hydrogen bond with water forms advantageous parameter for psychrophilicity (Jahandideh 2008). In antifreeze protein, the reoccurrence of Thr residues may block the growth of ice nucleation via the zipper like model, while, substitution of Thr with any other 12 amino acid types are found to be disadvantageous for the protein psychrophilicity (Chou 1992; Jahandideh et al. 2008). Our results are consistent with the above fact, that the occurrence of polar residues was more in psychrophiles which might be necessary for structure flexibility. Reports showed lower frequency of occurrence of Ser, Thr and Gln in thermophilic proteins than mesophilic proteins (Chakravarty and Varadarajan 2000; Kumar et al. 2000; Pack and Yoo 2004). These results coincides with our findings that thermophiles have lower occurrence of polar uncharged residues.

In GroESL protein, similar results were observed with charged residues and aromatic residues more in thermophiles than psychrophiles (Fig 7, Table 6). In GroES protein, lysine composition was seen more in thermophiles where as arginine and histidine was observed to be more in psychrophiles (Fig 7a). Frequency of tryptophan and cysteine occurrence was very low in thermophilic and psychrophilic GroES, therefore, it was not considered for the student t-test (Table 6). Polar uncharged amino acids were more in psychrophilic GroESL than thermophilic GroESL (Fig 7b).

Fig 7: Mean frequency of occurrence of each amino acid in GroES and GroEL proteins between thermophiles and psychrophiles.

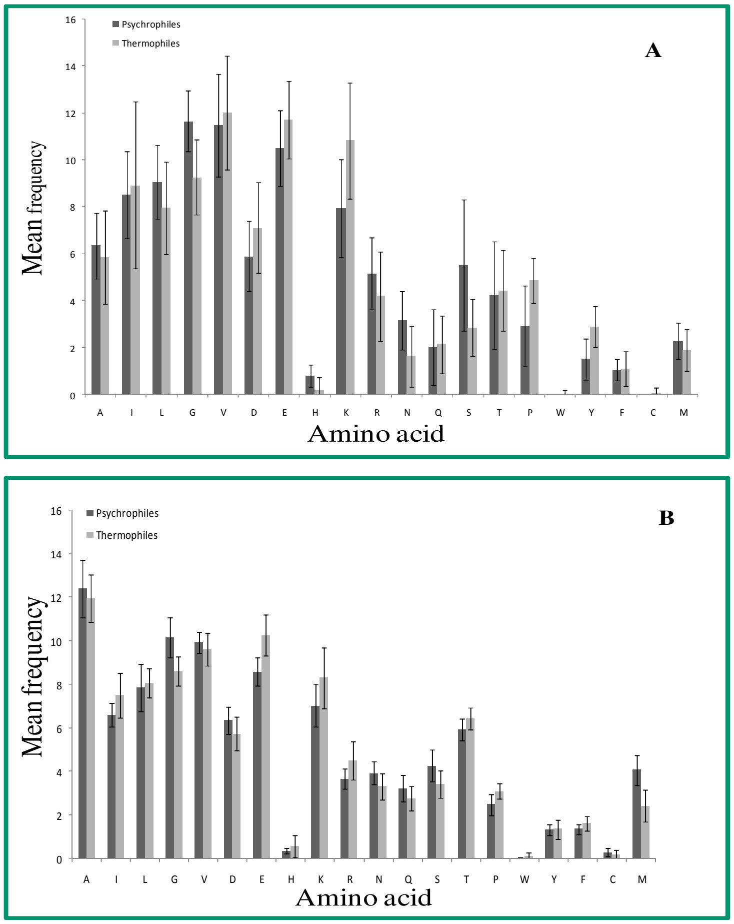


Fig 7: Mean frequency of occurrence of each amino acid in GroES and GroEL proteins between thermophiles and psychrophiles. Amino acid composition analysis of GroES (A) and GroEL (B) proteins shows significant differences in the composition of amino acids between thermophiles (grey bar) and psychrophiles (black bar). Amino acids are represented in single letter codes on the *X*-axis and the mean frequency of each amino acid occurrence is represented on the *Y*-axis.

Among the aliphatic amino acids, glycine is seen more in psychrophilic GroESL than thermophilic GroESL and is statistically more significant ($p < 0.0001$) (Table 6). Increased number of glycine residues serves to increase protein flexibility and stability by increasing the conformation entropy of the unfolded state in psychrophiles (Goldstein 2007; Jahandideh et al. 2008). In comparison with mesophiles, thermophiles showed fewer glycines in a particular region of the protein structure which might serve to decrease the flexibility of that region and thus, provide greater protein stability at high temperatures (Panasik et al. 2000). But another study showed no pattern of glycine occurrence in thermophilic proteins (Pack and Yoo 2004). Among the hydrophobic (aliphatic) amino acids, no differences were observed between the selected extremophiles, as these hydrophobic amino acids were known to be very important and significant in protein folding and refolding process in GroESL protein (Lin et al. 1995). It is reasonable to observe no significant differences in the overall content of hydrophobicity between thermophilic and psychrophilic GroESL protein (Table 6). The distribution of individual aliphatic amino acids was more than any other amino acids in conserved orthologs and GroESL protein which might be responsible for providing thermal stability to the protein core (Table 6). Compared to aliphatic amino acids, less frequently distributed amino acids were cysteine and tryptophan in both extremophiles (Table 5 and 6).

Table 6: Normalized frequency of each amino acid in the GroESL proteins between thermophiles and psychrophiles.

	Psychro GroES	Thermo GroES		Psychro GroEL	Thermo GroEL	
Amino acids	Mean±SD	Mean±SD	t-test (*p*-value)	Mean±SD	Mean±SD	t-test (*p*-value)
Ala	6.343±1.411	5.852±2.002	1.007 (0.3253)	12.410±1.303	11.973±1.075	1.086 (0.1849)
Ile	8.512±1.845	8.929±3.550	-0.585 (0.5627)	6.618±0.540	7.522±1.031	**-4.329 (0.0006)**
Lue	9.054±1.579	7.967±1.977	2.063 (0.053)	7.875±1.098	8.077±0.669	-0.615 (0.7425)
Gly	11.661±1.31	9.266±1.598	**5.515 (<0.0001)**	10.172±0.909	8.625±0.666	**5.576 (< 0.0001)**
Val	11.481±2.188	12.011±2.421	-0.747 (0.4656)	9.945±0.494	9.625±0.772	1.815 (0.0445)
Asp	5.894±1.502	7.104±1.933	-2.399 (0.0268)	6.361±0.633	5.736±0.784	2.961 (0.0054)
Glu	10.513±1.612	11.725±1.656	-2.353 (0.0317)	8.588±0.640	10.268±0.928	**-7.533 (< 0.0001)**
His	0.783±0.472	0.207±0.507	3.785 (0.0016)	0.358±0.137	0.574±0.493	-2.800 (0.0044)
Lys	7.932±2.093	10.824±2.490	**-4.193 (0.0005)**	7.044±0.986	8.304±1.411	-3.683 (0.0029)
Arg	5.156±1.538	4.188±1.913	1.890 (0.0742)	3.670±0.476	4.508±0.875	**-4.628 (< 0.0001)**
Asn	3.157±1.256	1.631±1.293	3.800 (0.0016)	3.945±0.542	3.321±0.592	3.552 (0.0023)
Gln	2.011±1.623	2.134±1.231	-0.249 (0.8074)	3.226±0.616	2.762±0.557	2.405 (0.0312)
Ser	5.503±2.795	2.854±1.226	3.218 (0.0082)	4.273±0.724	3.420±0.624	3.783 (0.0019)
Thr	4.231±2.279	4.451±1.721	-0.316 (0.7573)	5.947±0.509	6.431±0.512	-2.979 (0.0229)
Pro	2.923±1.715	4.869±0.959	-3.804 (0.0025)	2.484±0.471	3.105±0.361	-4.301 (0.0008)
Trp	-	-	-	0.019±0.064	0.124±0.152	-3.832 (0.0101)
Tyr	1.508±0.889	2.882±0.867	**-4.879 (0.0002)**	1.337±0.260	1.353±0.431	-0.170 (0.6348)
Phe	1.050±0.449	1.111±0.745	-0.374 (0.7112)	1.351±0.237	1.623±0.334	-3.318 (0.0018)
Cys	-	-	-	0.296±0.202	0.215±0.203	1.259 (0.727)
Met	2.287±0.768	1.886±0.899	1.592 (0.1287)	4.082±0.681	2.428±0.738	**7.512 (< 0.0001)**

Table 6: Normalized frequency of each amino acid in the GroESL proteins between thermophiles and psychrophiles. Frequency of each amino acid along with their standard deviation (SD) were mentioned. Student t-test shows the significant differences in amino acid composition between thermophiles and psychrophiles. Significant compositional differences are highlighted in bold between extremophiles. Dash infers calculation is not possible as frequency of occurrence of that amino acid is too low.

Temperature changes involves gradual changes in the protein sequences resulting in retaining the residues necessary for protein function and substituted by another residue which might be necessary for stability at extreme temperatures. Identification of the positions showing conserved differences in the amino acids can thus help in designing and producing the recombinant proteins with structural stability and enhanced activity at extreme temperature conditions.

3.2.4 To analyze the differences in amino acids at various positions of GroESL from either extremophiles to identify the amino acids that are responsible for their differential temperature optima:

GroES and GroEL protein sequences were aligned using T-coffee (http://www.tcoffee.org/) with default parameters. The aligned amino acids sequences were considered for amino acid frequency calculation at each position between thermophilic and psychrophilic GroES and GroEL proteins. From these comparisons, we have identified specific amino acid differences in GroES and GroEL sequences between thermophiles and psychrophiles which are shown in Table 7 and Table 8 respectively. Amino acids were grouped according to their side chain polarity, side chain charge and hydropathy index to understand the clear differences of effect of temperature on extremophiles that grow at different temperatures.

3.2.4.1 In thermo and psychro GroES:

In GroES protein sequence, it was observed that there is a temperature dependent change from a neutral to either negative or positive amino acids as the temperature increases (Table 7). Neutral amino acids at positions 29, 39, 43, 77 and 88 in psychrophilic GroES were replaced by charged amino acids in thermophilic GroES (Table 7). These uncharged to charged amino acid replacements between psychrophilic and thermophilic GroES sequences are highly conserved within each group of extremophilic bacteria. The polar uncharged amino acids at positions (10, 29, 77 and 88) in psychrophiles were found to be replaced by the charged amino acids in thermophiles (Table 7). Charged residues are involved in salt bridge formation in achieving stability of thermophilic protein in comparison to mesophilic proteins (Gromiha et al. 2008). Also hydrophobic amino acids at positions 39 and 43 in psychrophiles were replaced by basic amino acids in thermophiles (Table 7). Positively charged, histidine is conserved at position 15 of psychrophilic GroES (histidine 75%), whereas neutral hydrophilic amino acid (glycine 76%) was conserved in thermophilic GroES sequences respectively (Table. 7). According to Dayhoff and her colleagues, histidine to hydrophilic amino acids substitution is not a naturally acceptable point mutation (Conservative replacement) (Dayhoff et al. 1978). Even though, it is not an acceptable point mutation, histidine to glycine substitution from psychrophiles to thermophiles is the indication of molecular thermal adaptation (Table 7). In psychrophiles at positions 49, 59 and 62 charged amino acids were conserved, implying that these conserved sites might be necessary for the psychrophilic GroES protein activity at low temperature. Apart from these specific, significant differences between thermophilic and psychrophilic GroES sequences, few positions with completely variable amino acids such as 10, 26, 49, 59, 69 and 104 in thermophilic GroES and position 87 in psychrophilic GroES were observed (Table 7). These minor subtle differences in the primary sequence of extremophiles enable overall structural stability of GroESL protein, which is very much in accordance with the results of GroES-GroEL complex analysis between *E. coli* and *T. thermophiles* (Shimamura et al. (2004). Considerable conformational variation in the *cis* cavity and mobile loop of GroES-GroEL complex between *E. coli* and *T. thermophiles* was observed which could be due to the differences in protein sequences similarity (64 % for GroES and 71 % for GroEL between *E. coli* and *T. thermophiles*). The observed specific conserved amino acid substitutions in our analysis, supports the argument that during the course of adaptation to extreme conditions, selective changes in genetic material might have led to changes in amino acids for optimized function of protein. Such conserved differences in amino acids might have arisen during bacterial adaptation to extreme temperatures.

Table 7: Normalized frequency of amino acids at each position of aligned GroES protein sequence in thermophiles and psychrophiles.

Pos of AA in alignment	Psychro GroES	Frequency (%)	Thermo GroES	Frequency (%)
10	N,S,T (Phil)	91	**Variable**	NA
15	H,A (Pho)	91	Y,G,Q,N (Phil)	80
22	R (Base)	82	P,Y,T (Phil)	50
26	V,A (Pho)	46	**Variable**	NA
29	S,T (Phil)	91	E (Acid)	70
32	A (Pho)	100	**Variable**	NA
39	G (Phil)	73	D (Acid)	92
43	A (Pho)	55	K,R (Base)	98
49	R,K (Base)	73	**Variable**	NA
59	R,K (Base)	82	**Variable**	NA
69	P (Phil)	55	**Variable**	NA
77	S,Q,T (Phil)	64	R,K (Base)	78
82	E (Acid)	46	K,R (Base)	98
84	N,Y,G (Phil)	64	A (Pho)	62
87	**Variable**	NA	N,T (Phil)	98
88	T,S (Phil)	46	E,D (Acid)	96
96	V,F,L (Pho)	73	Y (Phil)	92
104	Y,S,N (Phil)	73	**Variable**	NA

Table 7: Normalized frequency of amino acids at each position of aligned GroES protein sequence in thermophiles and psychrophiles. Percentage of occurrence is mentioned in the column next to the amino acid. Well conserved amino acids among GroES proteins of both psychrophilic and thermophilic bacteria are excluded in this table. Phil, hydrophilic amino acids; Pho, hydrophobic amino acids; Base, negatively charged amino acids; Acid, positive charged amino acids; NA, not applicable. Variable represent no conserved amino acid is present in that column.

3.2.4.2 In thermo and psychro GroEL:

Amino acid substitutions were also compared between thermophilic and psychrophilic GroEL proteins. Unlike GroES subunit, GroEL subunit showed highly conserved amino acids throughout the protein length with few substitutions between the thermophiles and psychrophiles. We observed difference in many conserved residues of GroEL sequences between thermophiles and psychrophiles which are shown in Table 8.

Similar to GroES, GroEL sequence analysis showed temperature dependent changes from neutral to either positive or negative amino acid changes. At various positions of the multiple sequence alignment, hydrophilic amino acids in psychrophilic GroEL were replaced by charged amino acids in thermophilic GroEL at positions 43, 101, 136, 142, 158, 165, 183, 209, 212, 214, 219, 241, 269, 322, 346, 443, 444, 475, 476 and 497 (Table 8). Hydrophilic residues include polar uncharged residues which are suggested to increase the psychrophilicity of psychrophilic proteins (Jahandideh et al. 2008). Substitution of hydrophobic residues by polar residues in the cold active chitobiase showed improved interactions at low temperature (Lonhienne et al. 2001). Comparative studies on the metagenomic and genomic data from the psychrophilic organism showed that cold active enzymes contain relatively high levels of polar uncharged residues, which are necessary for the structural stability (Casanueva et al.

2010). Hydrophobic amino acids at positions (129, 144, 327, 343, 344, 456, 538 and 572) in psychrophilic GroEL were replaced by charged amino acids in thermophilic GroEL (Table 8). The percentage of charged residues is high in a set of proteins from hyper-thermophiles which are involved in thermal stability of proteins (Berezovsky and Shakhnovich 2005; Querol et al. 1996). These charged residues play a key role in the stabilization of proteins by ionic interactions at high temperature. High degree of hydrogen bond interactions increases thermal stability of proteins by increasing the stabilization energy (Tanner et al. 1996; Vogt and Argos 1997). Amino acid composition on the protein surface plays an important role in understanding the protein stability (Fukuchi et al. 2001). It is also interesting to note that, certain positions in psychrophiles contain charged residues, which are replaced by neutral hydrophilic (position 51, 65, 141, 186, 233, 244, 275, 342 and 441) or variable (position 5, 7, 11, 161,319, 337 and 467) amino acids in thermophile GroEL (Table 8). All these observed amino acid substitutions influence overall non-covalent interactions, which contribute to the stable structure of GroESL. Such conserved differences in amino acids might have arisen during bacterial adaptation to extreme temperatures.

Table 8: Normalized frequency of amino acids at each position of aligned GroEL protein sequence in thermophiles and psychrophiles.

Position of A.A	Psychro GroEL	Frequency (%)	Thermo GroEL	Frequency (%)
5	D,E (Acid)	91	**Variable**	NA
7	K (Base)	73	**Variable**	NA
9	G (Phil)	73	**Variable**	NA
11	D,E (Acid)	91	**Variable**	NA
21	N (Phil)	82	**Variable**	NA
43	S (Phil)	91	K (Base)	73
46	A (Pho)	73	T,S,G (Phil)	70
51	K (Base)	100	N (Phil)	66
65	K (Base)	82	P,Q,N,T (Phil)	64
81	A (Pho)	46	T (Phil)	99
101	N,T,Q (Phil)	55	R,K (Base)	70
106	A,L (Pho)	55	N,S (Phil)	80
115	D,E (Acid)	100	A,M,L,I,V,F (Pho)	70
129	A (Pho)	64	E,D (Acid)	82
135	S,T (Phil)	82	A (Pho)	46
136	T,Q (Phil)	64	K,R (Base)	73
141	D,E (Acid)	91	T,S,G,N (Phil)	71
142	T,N,Q,S (Phil)	100	R,K (Base)	79
143	K (Base)	82	D,E (Acid)	57
144	A (Pho)	73	E,D (Acid)	59
149	G (Phil)	82	A (Pho)	91
150	T,S (Phil)	100	**Variable**	NA
158	T,S (Phil)	46	E,D (Acid)	73
161	E,D (Acid)	64	**Variable**	NA
165	T,N,Q (Phil)	64	D,E (Acid)	97
183	Q,S (Phil)	73	K,R (Base)	82
186	E,D (Acid)	73	G,T,Q (Phil)	46
209	N,S (Phil)	46	D (Acid)	71
212	N,S,T (Phil)	46	K,R (Base)	86
214	S,T (Phil)	73	E (Acid)	93
216	D,E (Acid)	91	V,I,H (Pho)	80
219	S,N (Phil)	91	D,E (Acid)	71
221	F (Pho)	46	Y (Phil)	70
225	V,F (Pho)	91	Y,T,N (Phil)	79
233	R,K (Base)	91	T,Q,S,G (Phil)	50
241	Q,P,G (Phil)	64	K,R (Base)	68
244	K (Base)	64	Q (Phil)	77
245	A,M (Pho)	46	S,T,Q,Y (Phil)	73

Pos of AA in alignment	Psychro GroEL	Frequency (%)	Thermo GroEL	Frequency (%)
269	N (Phil)	73	K,R (Base)	90
273	I,A (Pho)	64	T (Phil)	77
275	K (Base)	91	S,T,N,Q (Phil)	82
311	E,D (Acid)	55	K,R (Base)	59
319	D (Acid)	64	**Variable**	NA
322	Q,T (Phil)	100	R,K (Base)	75
327	V,I (Pho)	46	R,K (Base)	51
337	D (Acid)	82	**Variable**	NA
342	E (Acid)	55	P,Q,S,T (Phil)	60
343	A (Pho)	46	E,D (Acid)	59
344	I,M,V,A (Pho)	46	D,E (Acid)	53
346	S,N,Q (Phil)	46	K,R (Base)	60
347	G,S,N (Phil)	64	A,M (Pho)	57
390	V,I,M (Pho)	82	T,P,S (Phil)	79
397	D (Acid)	73	H,L,A,M (Pho)	93
404	H (Pho)	82	S,N (Phil)	70
416	A (Pho)	64	P (Phil)	64
430	**Variable**	NA	E,D (Acid)	70
436	G,T (Phil)	73	**Variable**	NA
441	E,D (Acid)	73	G,P,N,T (Phil)	82
443	Q (Phil)	82	E,D (Acid)	77
444	T,N (Phil)	73	K,R (Base)	50
456	A (Pho)	82	E (Acid)	55
467	E,D (Acid)	73	**Variable**	NA
470	A,V (Pho)	55	G,P (Phil)	100
475	N (Phil)	82	E,D (Acid)	66
476	N,Q (Phil)	46	K,R (Base)	79
480	G (Phil)	91	**Variable**	NA
497	Variable	NA	E,D (Acid)	71
520	F (Pho)	64	N (Phil)	99
538	A,I,L,M,V (Pho)	73	K (Base)	64
540	**Variable**	NA	E,D (Acid)	93
566	G (Phil)	91	**Variable**	NA
567	M (Pho)	91	M,H,A,V,I,L (Pho)	68
568	G,P (Phil)	100	**Variable**	NA
571	G,P (Phil)	100	**Variable**	NA
572	M (Pho)	91	D,E (Acid)	48

Table 8: Normalized frequency of amino acids at each position of aligned GroEL protein sequence in thermophiles and psychrophiles. Percentage of occurrence is mentioned in the column next to the amino acid. Well conserved amino acids among GroEL proteins of both psychrophilic and thermophilic bacteria are excluded in this table. Phil, hydrophilic amino acids; Pho, hydrophobic amino acids; Base, negatively charged amino acids; Acid, positive charged amino acids; NA, not applicable. Variable represent no conserved amino acid is present in that column.

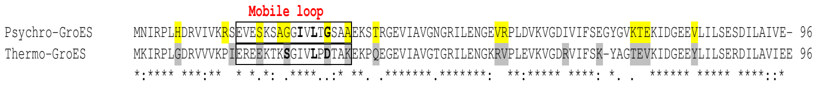
Positional conserved residue analysis showed that thermophiles preferred to have more charged amino acids while psychrophiles preferred to have more hydrophilic amino acids. This positional conserved residue information also provides much valuable information for identification of key residues that might contribute to thermal adaptability to the extreme temperature conditions. All these substitutions in the amino acid sequences have a direct effect the non-covalent interactions, which contribute towards the maintainance of the stable structure of GroESL. It implies that the stability results from multiple factors including improved hydrophobic package, optimized hydrogen bonding and favorable structural rearrangements.

3.2.5 Structural comparison of GroES between thermophiles and psychrophiles:

The amino acid changes observed between thermophilic and psychrophilic bacterial GroES and GroEL sequences might be necessary to sustain the interactions and retain the structural stability for optimal activity at extreme temperatures. Based on the differences that were observed in GroESL, we constructed consensus proteins of GroES and GroEL for thermophiles and psychrophiles (Fig. 8A and B).

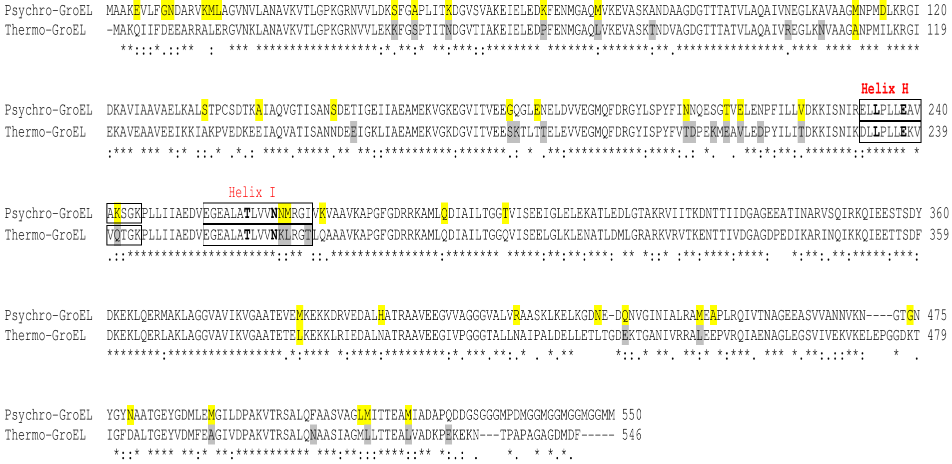
Fig 8: Sequence alignment of consensus GroESL protein sequences between thermophiles and psychrophiles.

(A) GroES



```
Psychro-GroES      MNIRPLHDRVIVKRSEVESKSAGGIVLTGSAAEKSTRGEVIAVGNGRILENGEVRPLDVKVGDIVIFSEGYGVKTEKIDGEEVLILSESDILAIVE- 96
Thermo-GroES       MKIRPLGDRVVVKPIEREEKTKSGIVLPDTAKEKPQEGEVIAVGTGRILENGKRVPLEVKVGDRVIFSK-YAGTEVKIDGEEYLILSERDILAVIEE 96
                   *:**** ***:**  * *.*: .****..:* **. .*******.*******:  **:***** ****: *. .  ****** ***** ****::*
```

(B) GroEL



```
Psychro-GroEL  MAAKEVLFGNDARVKMLAGVNVLANAVKVTLGPKGRNVVLDKSFGAPLITKDGVSVAKEIELEDKFENMGAQMVKEVASKANDAAGDGTTTATVLAQAIVNEGLKAVAAGMNPMDLKRGI 120
Thermo-GroEL   -MAKQIIFDEEARRALERGVNKLANAVKVTLGPKGRNVVLEKKFGSPTITNDGVTIAKEIELEDPFENMGAQLVKEVASKTNDVAGDGTTTATVLAQAIVREGLKNVAAGANPMILKRGI 119
                **:::*.::**  :  *** *****************:*.**:* **:***::******** *******:*******:**.****************.**** **** *** *****

Psychro-GroEL  DKAVIAAVAELKALSTPCSDTKAIAQVGTISANSDETIGEIIAEAMEKVGKEGVITVEEGQGLENELDVVEGMQFDRGYLSPYFINNQESGTVELENPFILLVDKKISNIRELLPLLEAV 240
Thermo-GroEL   EKAVEAAVEEIKKIAKPVEDKEEIAQVATISANNDEEIGKLIAEAMEKVGKDGVITVEESKTLTTELEVVEGMQFDRGYISPYFVTDPEKMEAVLEDPYILITDKKISNIKDLLPLLEKV 239
               :*** *** *:* ::.* .*.: ****.*****.** **::**********:*******.: * .**:***********:****:.: *.  . **:*:**:.*******::****** *

Psychro-GroEL  AKSGKPLLIIAEDVEGEALATLVVNNMRGIVKVAAVKAPGFGDRRKAMLQDIAILTGGTVISEEIGLELEKATLEDLGTAKRVIITKDNTTIIDGAGEEATINARVSQIRKQIEESTSDY 360
Thermo-GroEL   VQTGKPLLIIAEDVEGEALATLVVNKLRGTLQAAAVKAPGFGDRRKAMLQDIAILTGGQVISEELGLKLENATLDMLGRARKVRVTKENTTIVDGAGDPEDIKARINQIKKQIEETTSDF 359
               .::*********************::** ::.***************:******** *****:**:**:***: ** *::* :**:****:****:   *:**:.**:*****:***:

Psychro-GroEL  DKEKLQERMAKLAGGVAVIKVGAATEVEMKEKKDRVEDALHATRAAVEEGVVAGGGVALVRAASKLKELKGDNE-DQNVGINIALRAMEAPLRQIVTNAGEEASVVANNVKN----GTGN 475
Thermo-GroEL   DKEKLQERLAKLAGGVAVIKVGAATETELKEKKLRIEDALNATRAAVEEGIVPGGGTALLNAIPALDELLETLTGDEKTGANIVRRALEEPVRQIAENAGLEGSVIVEKVKELEPGGDKT 479
               ********:*****************.*:**** *:****:*********:*.***.**:.* . *.**      *::.* **. **:* *:***. *** *.**:.::**:    *  .

Psychro-GroEL  YGYNAATGEYGDMLEMGILDPAKVTRSALQFAASVAGLMITTEAMIADAPQDDGSGGGMPDMGGMGGMGGMGGMM 550
Thermo-GroEL   IGFDALTGEYVDMFEAGIVDPAKVTRSALQNAASIAGMLLTTEALVADKPEKEKN---TPAPAGAGDMDF----- 546
                *::* **** **:* **:*********** ***:**:::****::** *::: .    *.  * *.*.
```

Fig 8: Sequence alignment of consensus GroESL protein sequences between thermophiles and psychrophiles. Pair wise sequence alignment of GroES (A) and GroEL (B) consensus sequences of thermophile and psychrophile. Amino acids that differ between thermophilic and psychrophilic GroESL at a specific position, but conserved with in either thermophiles or psychrophiles are highlighted with yellow color. Based on the frequency of occurrence of each amino acid at a position, the consensus GroES and GroEL for thermo and psychrophiles were constructed.

Consensus sequences of GroES and GroEL for a group of 12 psychrophiles and 55 thermophiles were constructed by aligning protein sequences and by calculating the frequency of occurrence of residue at each position. The amino acid residue at each position in a consensus sequence was considered only when there was at least 50 % of occurrence at respective position. Consensus GroESL sequence for thermophiles and psychrophiles were generated and were designated as Thermo-GroES, Psychro-GroES, Thermo-GroEL and Psychro-GroEL (Fig. 8). There was 67 % sequence identity between consensus Thermo-GroESL and Psychro-GroESL sequences. In order to examine, whether conserved amino acid differences between psychro and thermo-GroESLs have any impact on structure and inter and intra-subunit interactions of GroESL complexes, which might be responsible for temperature dependency, 3D models of tri-subunit-GroESL complexes were constructed for Thermo-GroESL and Psychro-GroESL using InsightII software (Fig. 9). In nature, GroESL complex is a heptameric complex consisting of 7-GroES subunits and 7-GroEL subunits. In order to have an understanding of the inter-subunit and intra-subunit interactions, it is sufficient to build a tri-subunit complex, which consists of 3-GroES and 3-GroEL subunits. We achieved this using MODELER in InsightII. Interaction of GroES with large subunit GroEL stabilizes the ATP bound asymmetric GroESL complex and facilitates various interactions such as hydrophobic, ion-pair, hydrogen-bonding and vander-waals within the GroESL complex. All these interactions contribute to the strength of the GroESL complex by enhancing the functional stability of protein molecules (Fenton and Horwich 1997). The residues from ~16-32 called mobile loop structure of the GroES subunit is involved in such interactions with GroEL *cis* ring by stabilizing the complex structure of GroESL (Shewmaker et al. 2001). We identified four conserved differences in this loop region between psychrophiles and thermophiles; in psychrophile, alanine at position 22 is replaced by lysine, glycine at position 23 and 29 by serine and aspartic acid respectively (Fig. 9 A-C). The conformational differences in the mobile loop are clearly seen in the 3D models of Psychro-GroES and Thermo-GroES. Aspartic acid at position 29 of Thermo-GroES forms 11 hydrogen bond interactions, whereas glycine at the same position in psychro-GroES, showed only 7 interactions in tri-GroESL subunit complex (Fig. 9 A-C). Overall, Thermo-GroES has more interactions in the mobile loop region with Thermo-GroEL in the Thermo-GroESL trisubunit complex, as compared to the Psycrho-GroES with Psychro-GroEL in the Psychro-GroESL trisubunit complex (Fig. 9 A-C). The conserved amino acids in the mobile loop of Thermo-GroES when compared to Psychro-GroES, shows more numbers of intra-molecular interactions within the mobile loop may stabilize the loop structure at high temperatures.

Fig 9: Schematic representation of GroES interaction with GroEL in tri-subunit GroESL complexes of Thermo-GroESL and Psychro-GroESL.

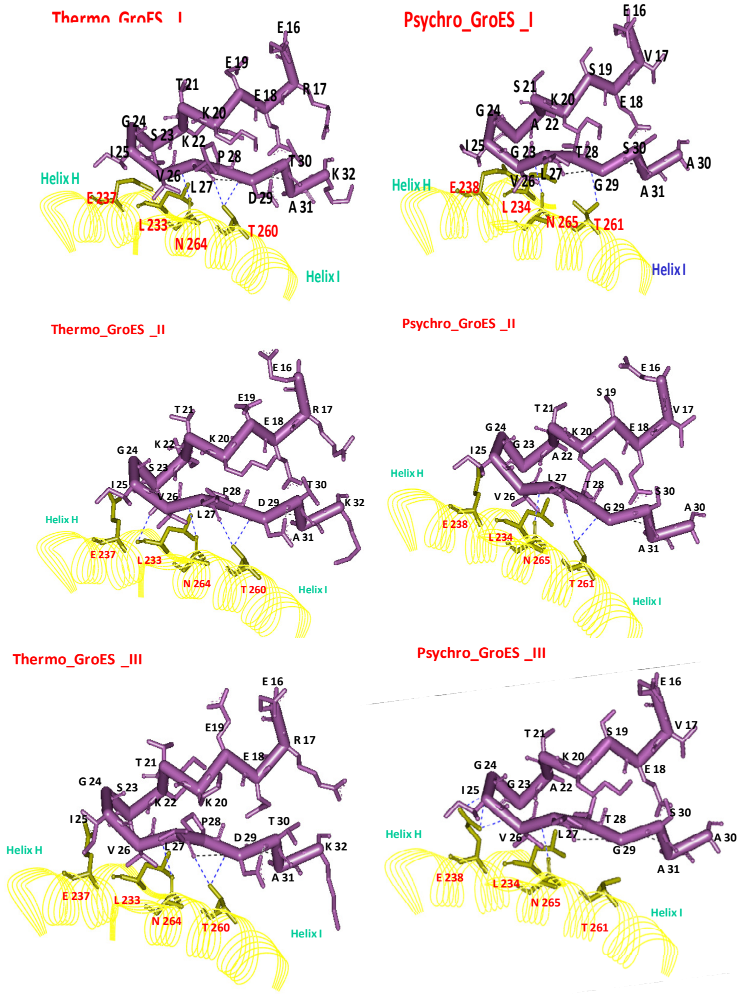


Fig 9: Schematic representation of GroES interaction with GroEL in tri-subunit GroESL complexes of Thermo-GroESL and Psychro-GroESL. Hydrogen bond interactions are shown in dotted lines. Subunit I (A), II (B) and III (C) represents the mobile loop regions of all three sub GroESL complexes of tri-subunit GroESL complex. Amino acids in red color font represent the conserved amino acid of GroEL. Aspartic acid shows more

number of interactions in Thermo-GroESL complex, than Glycine in Psychro-GroESL complex.

The intra and inter-molecular interaction made by amino acids in the mobile loop of GroES, when compared to other GroES subunits of the same tri-subunit complex are asymmetric (Shimamura et al. 2004). Aspartic acid at position 29 of Thermo-GroES, forms 4, 3 and 4 interactions in subunit I, II and III respectively (Fig 9 A-C). We suggest that this variation in the mobile loop amino acids and their interactions may be required for maintaining the compactness, thereby, structural stability to retain the activity at various temperatures. These specific amino acid alterations in the mobile loop of thermo and psychro-GroES may be required for maintaining temperature dependent interactions to strengthen the complex structure in either extremophiles.

On the other hand, it was observed that all the residues involved in catalysis are conserved in GroESL sequences of the two extremophiles. Amino acids crucial for binding to ATP (V31, G32, P33, T91, I150, I454, G415, N479, A480, A481, I493, and D495) (Boisvert et al. 1996) are well conserved in GroEL sequences of the two extremophiles. Aspartic acid at position, 398 involved in ATP hydrolysis and in progression of the chaperonin cycle, is conserved in all the GroEL sequences. Substrate binding residues Y199, Y203, P204, L234, L237, L259, V263 and V264 are also well conserved in all psychrophilic and thermophilic GroELs indicating the importance of these amino acids in the refolding activity of GroEL (Coyle et al. 1997; Fenton et al. 1994). Replacement of C-terminal non-conserved end of *E. coli* GroEL with psychrophilic C-terminal region, replacement of thermophilic and psychrophilic C-terminal region with *E. coli* C-terminal region and subsequent temperature dependent refolding assays suggested that C-terminal region, 503-548 amino acid residues was necessary for temperature dependent reaction (Nakamura et al. 2004). Site directed mutagenesis of *E. coli* GroEL, replacing leucine at 524 with isoleucine in the C-terminal region indicated that presence of selected amino acid within this region is responsible for temperature dependent activity (Nakamura et al. 2004). However, conserved differences observed in the present work, are spread throughout the GroESL protein sequences between psychrophies and thermophiles. It is possible that, not only the C-terminal region, but also all those conserved and selective changes in the GroESLs might be responsible for temperature dependent optimization of protein function. Comparative analysis of structure and sequence of GroESL between thermophiles and psychrophiles showed that during adaptation to diverse temperature conditions, genes undergo selective changes to retain the function of a protein. Thus, identified conserved amino acid changes between psychrophile and thermophile GroES and GroELs are assumed to be very important for their thermal adaptability.

3.2.6 Characterization of consensus synthetic thermo-GroESL and psychro-GroESL:

Sequence comparison by *insilico* methods of molecular chaperones (GroEL and GroES) from thermophiles and psychrophiles had identified the key residues that might be necessary for thermal adaptation. To understand the importance of these key residue differences we have constructed the consensus protein sequence derived from alignment of protein sequences of GroESL of 55 thermophiles and 12 psychrophiles. The significance of the conserved amino acid differences was analyzed by protein denaturation studies (thermal denaturation) of synthetic protein of GroESL of thermophile and psychrophile. Consensus sequences of GroES and GroEL for a group of 12 psychrophiles and 55 thermophiles were constructed. The amino acid residue at each position in a consensus sequence was considered only when there was at least 50 % of occurrence at respective position. These consensus GroESL proteins were reverse translated using the *E. coli* genetic code and the synthetic psychro-*groESL* and thermo-*groESL* were synthesized by GenScript USA Inc. (Piscataway, New Jersey, USA). The synthesized *groESL* genes of thermophile and psychrophile were cloned into pUC57 vector separately by GeneScript with the mentioned restricted enzyme sites so that synthetic genes can be excise from the pUC57 vector (Fig 3).

3.2.7 Expression and purification of synthetic GroESL protein:

Synthetic *groESL* genes of the thermophiles and psychrophiles were cloned in to pET28a(+).

Fig 10: Expression of synthetic thermophilic and psychrophilic GroEL.

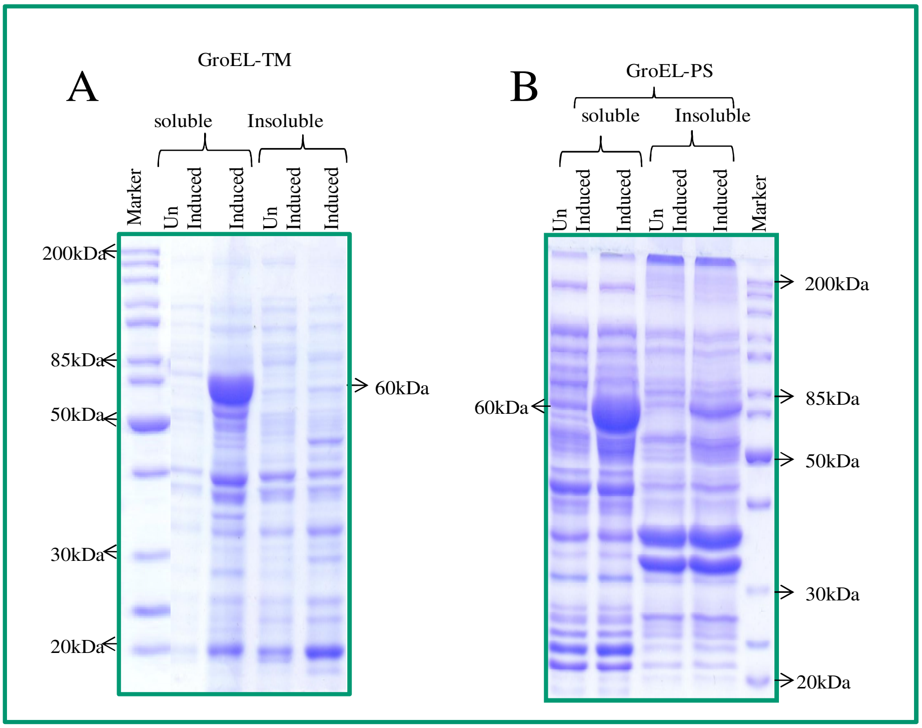


Fig 10: Expression of synthetic thermophilic and psychrophilic GroEL. Soluble and insoluble fraction of uninduced and induced samples of synthetic thermo-GroEL (A) and psychro-GroEL (B) protein expressed in *E. coli* (*Rosetta* stain) after induction with 0.4 mM IPTG. (A) Expressed synthetic thermo GroEL protein was resolved on SDS gel (15%) to approximately at 60 kDa (lane 3) in the soluble fraction; Marker, 200kDa size. (B) Expressed synthetic psychro GroEL protein was resolved on SDS gel (15%) to approximately at 60 kDa (lane 2) in the soluble fraction; Marker, 200 kDa size.

Fig 11: Expression of synthetic thermophilic and psychrophilic GroES.

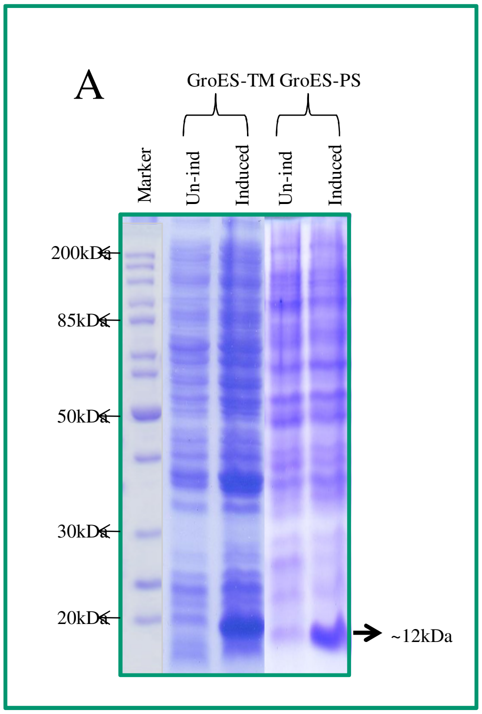


Fig 11: Expression of synthetic thermophilic and psychrophilic GroES. Uninduced and induced samples of synthetic thermo and psychro-GroES protein expressed in *E. coli* (*Rosetta* stain) after induction with 0.4 mM IPTG. (A) Expressed synthetic thermo (lane 3) and psychro (lane 5) GroES protein was resolved on SDS gel (15%) to approximately at 12 kDa in the induced sample; Marker, 200 kDa size.

Synthetic *groESL* gene of thermophile and psychrophile was expressed in *Roseta* strain of *E. coli* and the protein was resolved on the SDS PAGE. GroESL proteins were expressed in the soluble fraction for both thermophile and psychrophile. Synthetic thermo and psychro GroEL protein was expressed with the band size of 60 kDa (Fig 10) and GroES protein with the band size of 10 kDa (Fig 11). The samples were purified using Nickel Affinity column following the procedure as explained in the methodology. The purified elutes of thermo and psychrophiles GroESL were run on the SDS gel and found single band of approximately of 60 kDa for GroEL protein and 12 kDa for GroES protein in both thermophile and psychrophile (Fig 12and 13).

Fig 12: Purification of synthetic thermophilic and psychrophilic GroEL protein using Ni-NTA column.

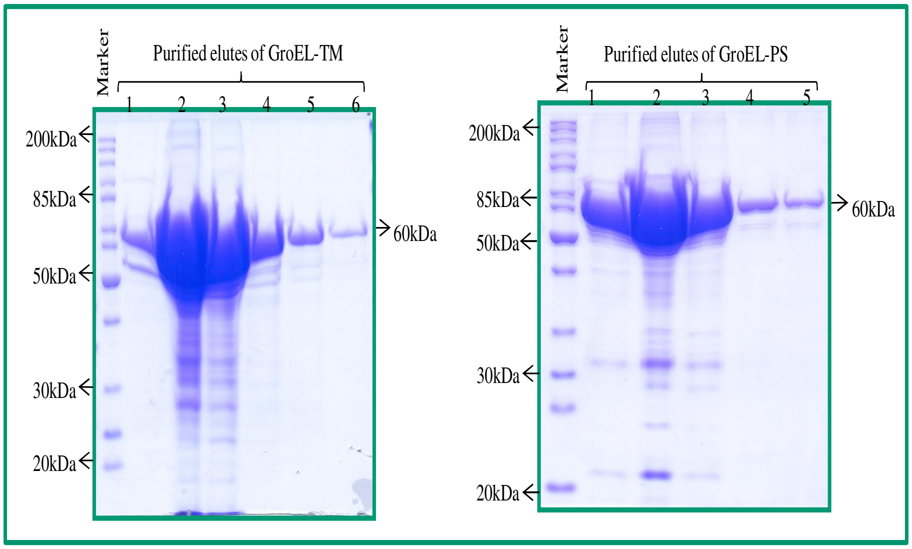


Fig 12: Purification of synthetic thermophilic and psychrophilic GroEL protein using Ni-NTA column. Synthetic thermo and psychro-GroEL protein are expressed in *E. coli* and purified using Ni-NTA column. Purified elutes of induced synthetic thermo-GroEL and psychro-GroEL proteins showed a required size of 60 kDa. Lane numbers 4, 5 and 6 for thermo-GroEL and Lane numbers 3, 4 and 5 for psychro-GroEL were pooled separately and concentration was checked. Marker, 200 kDa size.

Samples numbers 4, 5 and 6 of GroEL thermophile and 3, 4, 5, 6 and 7 of GroEL psychrophile were pooled separately and concentrated for thermal denaturation studies.

Fig 13: Purification of synthetic thermophilic and psychrophilic GroES protein using Ni-NTA column.

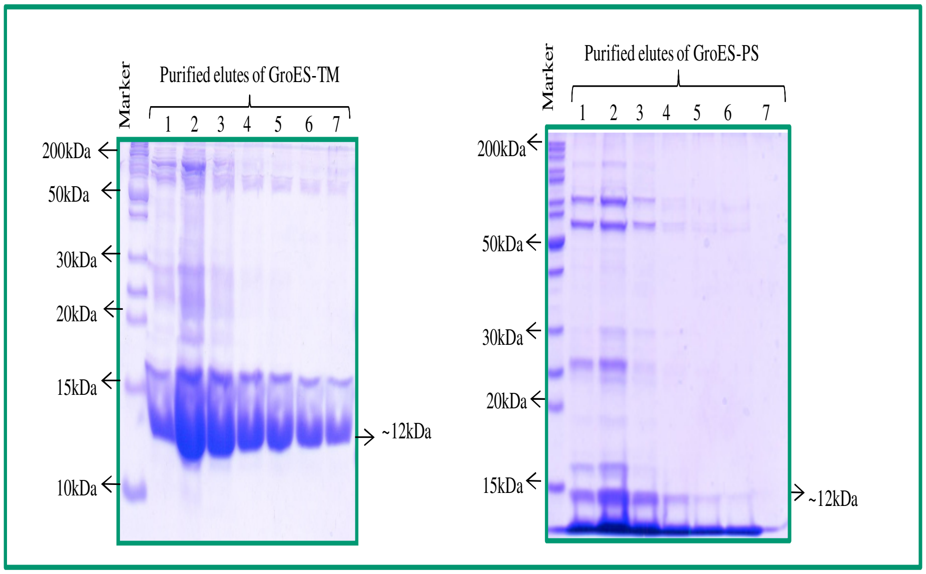


Fig 13: Purification of synthetic thermophilic and psychrophilic GroES protein using Ni-NTA column. Synthetic thermo and psychro-GroES protein are expressed in *E. coli* and purified using Ni-NTA column. Purified elutes of induced synthetic thermo-GroES and psychro-GroES proteins showed a required size of ~12 kDa. Lane numbers 3, 4, 5, 6 and 7 for thermo-GroEL and Lane numbers 4, 5 and 6 for psychro- GroEL were pooled separately and concentration was checked. Marker, 200 kDa size.

3.2.8 Thermo-GroEL is relatively more stable than psychro-GroEL:

Thermal denaturation of consensus thermo-GroEL and psychro-GroEL showed a cooperative loss of secondary structure as temperature is increased (Fig 14).

Fig 14: Thermal stability of purified thermo-GroEL and psychro-GroEL proteins.

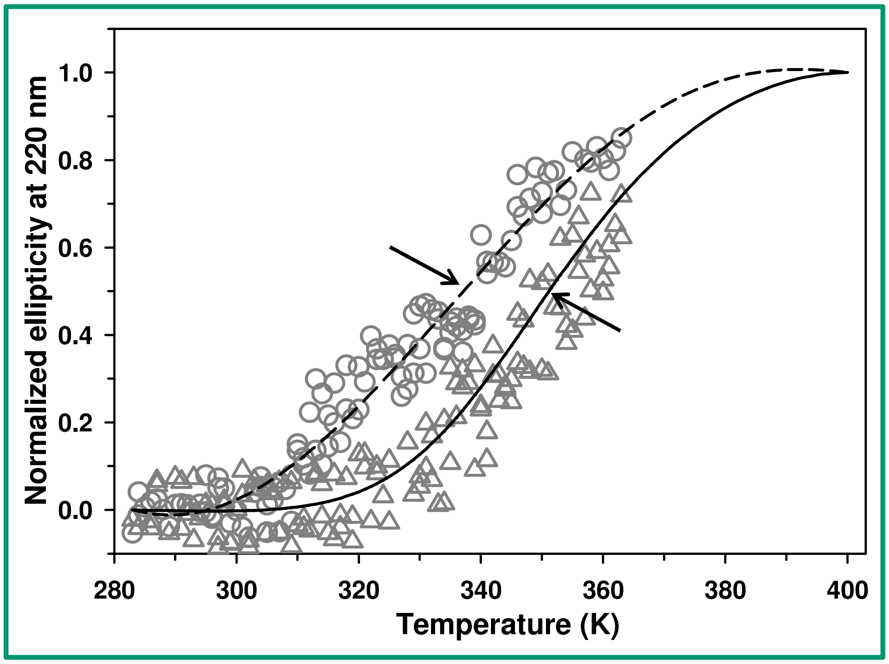


Fig 14: Thermal stability of purified thermo-GroEL and psychro-GroEL proteins. Structural changes were followed by monitoring ellipticity at 220 nm. Open circles (o) psychro-GroEL, open triangles (Δ) thermo-GroEL. The data was extrapolated using Gibbs-Helmholtz equation. The arrows indicate the temperature at which 50 % of the structure of the protein is lost (T_m).

However, the structure was not completely lost at the highest temperature practically attainable. Therefore, the obtained data were extrapolated for complete denaturation and the temperature midpoints (T_m) were calculated. It was found that psychro-GroEL and thermo-GroEL proteins showed the T_m value of 64 °C and 79 °C, respectively (Fig 14). The significant temperature differences confirm the fact that these two proteins have clearly a distinct thermal stability. Further insight on the data shows that up to nearly 55 °C, the thermo protein retains 10 % of its secondary structure where as the psychro protein lost approximately one third of its structure at the same temperature (Fig 14). These data corroborate with the idea that individual conserved amino acids within each group of thermophilic or psychrophilic GroELs play a key role in their thermal stability and adaptability.

3.3 Conclusion:

Comparative study between thermophiles and psychrophiles showed that there are differences in the amino acid composition. With charged (D, E, K, R), aromatic (F, Y) and proline residues being more in thermophiles and polar amino acids (S, T, Q, N, C, M) are more in psychrophiles. Aliphatic amino acids were same in both groups, which might be necessary to maintain overall hydrophobicity. These differences in the residue composition might be necessary for the thermal stability of the protein molecule. Positional conserved amino acid analysis showed that many hydrophilic and hydrophobic amino acids in psychrophilic GroESL were replaced by charged group in thermophilic GroESL with increase in temperature. Positional conserved amino acid differences of GroESL between thermophiles and psychrophiles might have arisen during bacterial adaptation to extreme temperatures and also forms the reasons for their different temperature optima. Thermal denaturation studies of the synthetic proteins of thermo-GroEL shows relatively more stablility than psychro-GroEL infers that key residues identified might be necessary for providing structural stability.

Chapter 4:

This chapter presents how selective the thermophiles and psychrophiles are in using different codons for optimizing their GC content of genome with respect to different temperature extremes. It also describes the conservative and non-conservative substitutions that might have necessarily happened for survival of these thermophiles and psychrophiles to extreme temperatures during the course of adaptation. Over all, thermophiles evolved with relatively high GC content genomes when compared to psychrophiles. Non-synonymous substitutions were predominant in thermophilic genomes when compared to psychrophiles, implying habitat dependent negative selection might have taken place during their evolution as extreme thermophiles. Rate of synonymous substitutions was almost similar in either extremophiles. Nature of nucleotide and amino acid substitutions that might have occurred during the course of adaptation to extreme temperatures for these extremophiles is discussed in detail.

Codon usage and synonymous and non-synonymous substitution rate analysis of conserved proteins between the selected thermophiles and psychrophiles were performed in this present chapter.

4.1 Introduction:

Genome sequence information is available for the large number of prokaryotic and eukaryotic organisms, which provides huge amount of valuable datasets for understanding similarities and differences between genomes of different organisms. Comparative genome analysis of multiple organisms provides answer to several questions related to evolution, adaptation to environmental niche, disease cause, special traits etc. There are many parameters for understanding the DNA sequence evolution such as mutations, genome composition, G+C composition, codon usage, synonymous and non synonymous substitution rate, effective number of codon (Nc) (Wright 1990), codon adaptation index (CAI) (Sharp and Li 1987) and frequency of optimal codon (F_{op}). According to Tamura (1992), for estimating the number of nucleotide substitutions per site between two sequences, two factors are important to consider, which include the inequality of the rates of transitional transversional and G+C content bias (Tamura 1992). It is important to analyze the substitutions at nucleotide and amino acid level that might have necessarily occurred during the course of evolution of bacteria as extreme thermophiles and psychrophiles. For this, codon usage bias and synonymous and non-synonymous substitution between thermophiles and psychrophiles was examined.

4.1.1 Codon profiling:

The biological information of a living organism is stored in the long sequence of nucleic acid containing four letters A, T, G and C. Group of three consecutive bases in the coding sequence is called a codon, which encodes an amino acid or to a signal that terminates translation. The mapping from codons to amino acids is called genetic code. There are total 64 codons (four bases for the third position in the codon: 4^3) that encode twenty different amino acids. There is more than one codon that encodes same amino acid, which is called degeneracy or redundancy of the genetic codon. Codons that encode same amino acid are called synonymous codons. Most of synonymous codons differ by only one base in their 3rd position (wobble position). The genetic code is conserved in many organisms but the usage of codons differs between the organisms (Ermolaeva 2001). Codon profiling is performed to understand the nucleotide compositional variations of the coding DNA sequences between thermophiles and psychrophiles. Frequency of one of the several codons is found more than expected by chance. Such preference of genetic code is referred as codon usage. Codon usage bias is defined as the differences in the frequency of occurrence of codons in the coding DNA. Codon usage analysis between thermophiles and psychrophiles was investigated in our present study as the strength of codon bias varies across the genes between and within the organisms (Grantham et al. 1980a; Gouy and Gautier 1982; Ikemura 1985). Strength of codon bias also varies between the orgnanism. In some organisms codon bias is very strong (strength) while in others, it was used with similar frequencies (Hershberg and Petrov 2008). Codon bias was vaies between the genes, highly expressed genes show greater levels of bias than and weakly expressed genes (Gouy and Gauiter 1982; Ikemura 1985). In *E. coli,* highly expressed genes showed preferred set of synonymous codons which are recognized by the abundant tRNAs, while weakly expressed genes did not show any bias in codon usage (Andersson and Kurland 1990). Many statistical approaches are available for the analysis of the codon usage, which includes prediction of the Relative Synonymous Codon Usage (RSCU), Codon Adaptation Index (CAI), Relative Codon Adaptation (RCA) and effective number of codons (Nc) (Sharp and Li 1986; Sharp and Li 1987; Fox and Erill 2010; Wright 1990). Codon usage was first examined in *E. coli* and *S. cerevisiae* where highly expressed genes showed bias towards particular subset of codons (Sharp et al. 1988). Codon bias was also well reported in the unicellular prokaryotes to multicellular eukaryotes and even in the Drosophila (Gouy and Gauiter 1982; Akashi 1994). Ingvarsson, (2008), reported that five closely related species of Populus (*P. tremula, P. deltoids, P. euphratica, P. trichocarpa* and *P. nigra* showed difference in synonymous codon usage (Ingvarsson 2008). According to Grantham et al. (1980a), each genome has its own characteristic patterns of synonymous codon usage and all genes of the genome have the similar choice of synonymous codon usage. These finding have been designated as “genome hypothesis” (Grantham et al.1980a). Many computational tools are available for understanding the codon usage which includes CodonW (http://codonw.sourceforge.net/), GCUA (General Codon Usage Analysis) (http://bioinf.may.ie/GCUA/), INCA (INteractive Codon Analysis software) (http://bioinfo.hr/research/inca/), ACUA (Automated Codon Usage Analysis tool) (http://www.bioinsilico.com/acua), PAL2NAL (http://www.bork.embl.de/pal2nal/), CALcal (http://genomes.urv.cat/CAIcal/E-CAI/) and others.

4.1.2 Synonymous and non-synonymous substitution:

Although mutation and substitution are often used interchangeably, there is a subtle but important difference. A nucleotide mutation is a base change (whether synonymous or non-synonymous) such that the mutant and wild-type forms co-exist in a population. A nucleotide substitution is a base change between two populations. Thus, a mutation only becomes a substitution when the most recent common ancestor of the entire population carried that mutation. Synonymous substitution or silent substitution is a substitution of one base within the triplet codon, which does not modify the amino acid. For example, codon GCT encodes alanine amino acid, if T is changed to A then the codon becomes GCA which also encodes alanine. Synonymous codons usually differ in the single nucleotide at the third position of a codon, but in some amino acids it differs in the second position also. Non-synonymous substitution which results in a change of amino acid which are of three types conservative, semi-conservative and radical (http://en.wikipedia.org/wiki/Synonymous_substitution). Conservative substitution is substitution from one amino acid to another amino acid but having same physiochemical properties. Semi conservative substitution is substitution from positive to negative charged amino acid and radical substitution is substitution resulting in entirely different amino acid. Synonymous and non-synonymous substitution rate analysis was used as tool for estimating the evolutionary distances of closely related species (Kafatos et al. 1977; Kimura 1977; Perler et al. 1980; Miyata and Yusunaga 1980).

Many mathematical methods were developed to estimate the nucleotide synonymous and non-synonymous rate between the DNA sequences (Perler et al. 1980; Miyata and Yusunaga 1980; Li et al. 1985; Nei and Gojobori 1986; Pamilo and Bianchi 1993; Goldman and Yang 1994; Muse and Gaut 1994; Comeron 1995; Ina 1995). According to Perler's method, equal weights were given to two or more possible evolutionary pathways between a pair of codons (i.e., for one codon to be transferred to another codon was considered as evolutionary pathway). According to the Miyata and Yusunaga (1980) and Li et al.'s (1985) methods, larger weight was given to an evolutionary pathway involving silent substitution and lower weight was given to a pathway involving amino acid altering substitution. Nei and Gojobori (1896), gave equal weights to all evolutionary pathways for given pair of codons in their method (Nei and Gojobori 1986). The results obtained by all four methods have showed approximately more or less the same results (Nei and Gojobori 1986)

Most of the tools calculate number of synonymous and non-synonymous substitutions per site by a method described by Nei and Gojobori (Nei and Gojobori 1986). A nucleotide site is classified as non-degenerate, two fold degenerate, three fold degenerate and four fold degenerate, depending on how often nucleotide substitutions will result in amino acid replacement (Li et al. 1985). A nucleotide position in a codon is said to be non-degenerate site, when replacement of the nucleotide at that position by any other three nucleotides results in change of the amino acid coded by that codon. The rate of nucleotide substitution is low at the non-degenerate sites. A position of a codon is said to be a twofold degenerate site when replacement of two out of four nucleotides at that position results in same amino acid. For example, GAA is one of the glutamic acid codons. The third position is a two fold degenerate site, because substitution of A at position three by G (transition change) specifies glutamic acid. There is only one threefold degenerate site where changing to three of the four nucleotides may have no effect on the amino acid but changing to the fourth possible nucleotide, thus results, in change in amino acid substitution. For example, third position of isoleucine codons AUU, AUC, and AUA are three fold degenerate sites. All three encode for isoleucine, change in the nucleotide at third position with guanine results in AUG which encodes a different amino acid, methionine. A position of a codon is said to be a fourfold degenerate site when substitution of nucleotide by any nucleotide at this position specifies the same amino acid. For example codons GGA, GGG, GGC and GGU encode the amino acid, glycine. According to the Nei and Gojobori, synonymous (S) and non-synonymous (N) sites are scored based on whether the site is a two fold, three fold or a four fold degenerate site, when comparing two nucleotide sequences. S = 1 and N = 0 for a four fold degenerate site. For a three fold degenerate site, S = 2/3 and N =1/3, for a two fold degenerate site S= 1/3 and N = 2/3 and for a non degenerative site S=0 and N=1.

According to the Nei and Gojobori, two sequences are compared by considering the aligned codons and count the number of synonymous and non-synonymous nucleotide differences for each pair of codons. For example, when codon GTT (Val) is compared to GTA (Val) codon, there is only one nucleotide difference between the codons i.e. one syonymous subtitution. For the two nucleotide differences between the codons, for example, TTT is compared to GTA there are two possible ways to obtain the differences. 1) TTT (Phe) ↔ GTT (Val) ↔ GTA (Val) and 2) TTT (Phe) ↔ TTA (Leu) ↔ GTA (Val). First way involves one synonymous and one non-synonymous substitution and second way involves two non-synonymous substitutions. The s_d and n_d will be 0.5 and 1.5 respectively. For the three nucleotide differences between the codons there will be six different possible pathways between the codons. S_d and n_d values are calcualted in the same way as above. Total number of synonymous and non-synonymous differences can be obtained by summing up these values (Nei and Gojobori 1986). Number of synonymous substitutions (d_S) and non-synonymous substitutions (d_N) per site was calculated by the formula developed by Jukes and Cantor (1969)

$$d = -\frac{3}{4}\log_e(1 - \frac{4}{3}p)$$

where p is either ps or PN.

If d_N/d_S ratio >1 signifies the adaptive selection and ratio <1 signifies purifying selection (Akashi 1999; Crandall et al. 1999; Sharp 1997).

Many computational tools available for the calculation of d_N/d_S ratio, which includes SNAP (Synonymous Non-synonymous Analysis Program), WINA (Window Analysis), SWAKK (Sliding Window Analysis of Ka and Ks), PAL2NAL (v12), DNaSP (DNA Sequence Polymorphism), IDEA (Interactive Display For Evolutionary Analyses) and others.

4.2 Results and Discussion:

Guanine and cytocine are an important attributes to understand the gene evolution of bacteria. The average GC content of thermophilic and psychrophilic genomes is 41.83 % and 53.36 % respectively (Fig 15). This suggests that, GC content of genomes of psychrophilic bacteria is relatively low when compared to the genomes of thermophiles. Thermophilic bacteria evolved with relatively higher GC content when compared to psychrophilic bacteria. It is because G is linked to C by three stable hydrogen bonds whereas only two hydrogen bonds link A to T. Therefore, the energy required to break the bonds between GC is more than required to break the bonds between AT. High GC content in thermophiles is necessary to protect the DNA and RNA from high temperature denaturation (Forterre, in press). According to Galtier and Lobry (1997), there is no co-relation between the GC content of the DNA and the optimal growth temperature of an organism (Galtier and Lobry 1997). However, thermophiles and hyper-thermophiles have higher GC content compared to mesophiles. GC content is one of the major factors that influence DNA, mRNA and tRNA stability (Fujita and Kanehisa 2005). Presence of more number of purine nucleotides prevents aggregation of mRNA molecules in thermophiles (Lao and Forsdyke 2000).

Fig 15: GC content of genomes of thermophilic and psychrophilic bacteria.

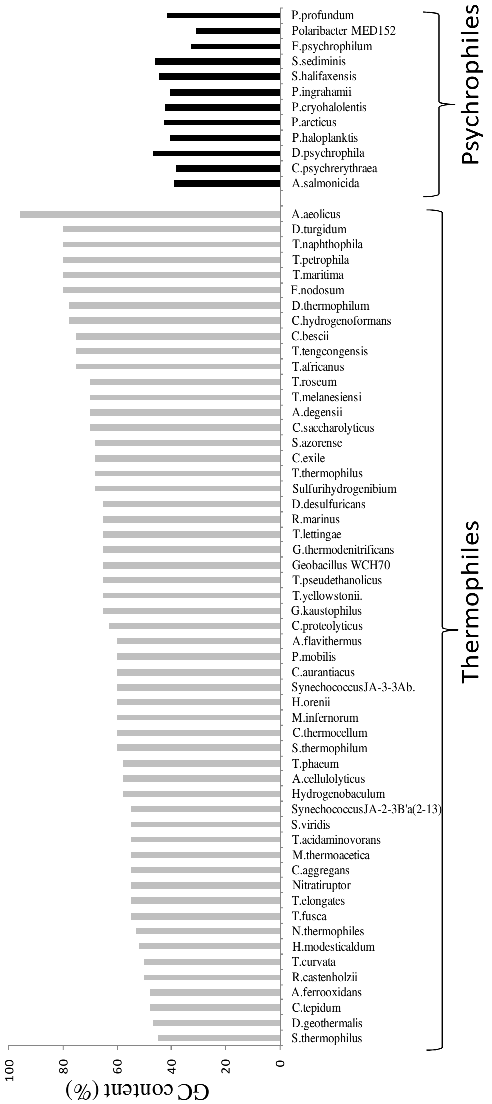


Fig 15: GC content of genomes of thermophilic and psychrophilic bacteria. The average GC content of thermophilic and psychrophilic genomes is 41.83 % and 53.36 % respectively. GC content of genomes of psychrophilic bacteria is relatively low when compared to the genomes of thermophiles. Grey color bar represent thermophiles and black bar represent psychrophiles.

Large variation was observed in the GC content within thermophilic genomes as compared to psychrophilic genomes, could be due to high temperature effect on the genomic DNA (Fig 15) (Kreil and Ouzounis 2001). Less variation in GC content of psychrophiles could also infer the low mutational rate in their genomes (Basak and Ghosh 2006). Kreil and Ouzounis (2001), has demonstrated that thermophiles have higher GC content than mesophiles, which are necessary for protection of DNA from extreme temperatures (Kreil and Ouzounis 2001).

To investigate the codon usage between thermophiles and psychrophiles, the orthologous proteins, which are conserved in all the selected thermophiles and psychrophiles were used.

4.2.1 To analyze the codon usage between thermophilic and psychrophilic conserved proteins to understand the evolution of thermophiles with high GC content genomes.

G+C content is one of the important factors, which was considered to measure the codon usage bias among the genes. According to Chen et al. (2004), pattern of codon usage depends on two parameters one being the G+C content and other being nucleotide bias. In order to check how selective the psychrophiles and thermophiles are in using different codons for optimizing their GC content of genome, codon usage bias was analyzed. We considered third position of the triplet codon (wobble position) which is a four fold degenerate site i.e., change in the nucleotide in the third position mostly leads to no change in amino acid. First and second positions of the codon are non-synonymous sites and have strong purifying selection where as, third position in codon is synonymous and have weak purifying selection (Basak et al. 2004). Codons are divided into G/C and A/T ending codons based on the presence of G/C or A/T in the third position. Therefore, it is interesting to see whether thermophiles maintain GC ending codons at wobble position as compared to psychrophiles. Codon profiling is performed to see whether codon usage bias is the basis for the thermophiles and psychrophiles to maintain the high and low GC content of their genome respectively. Open reading frames were retrieved for all 119 conserved orthologs of thermophiles and psychrophiles using *Perl* script and submitted to another *Perl* script which calculated the frequency of occurrence of A/T and G/C ending codons for each amino acid. Mean frequency of A/T and G/C ending codons of each amino acid of conserved proteins between thermophiles and psychrophiles were compared.

Codon usage analysis of conserved proteins showed that codons ending with A/T were preferred mostly by psychrophiles where as G/C ending codons by thermophiles (Fig 16, Table 9). The result obtained with the Lynn et al. (2002), showed that thermophiles preferred to have high GC ending codons and while mesophiles preferred to have more AT ending codons (Lynn et al. 2002).

Fig 16: Mean frequency of G/C and A/T ending codons of each amino acid in the conserved genes between thermophiles and psychrophiles.

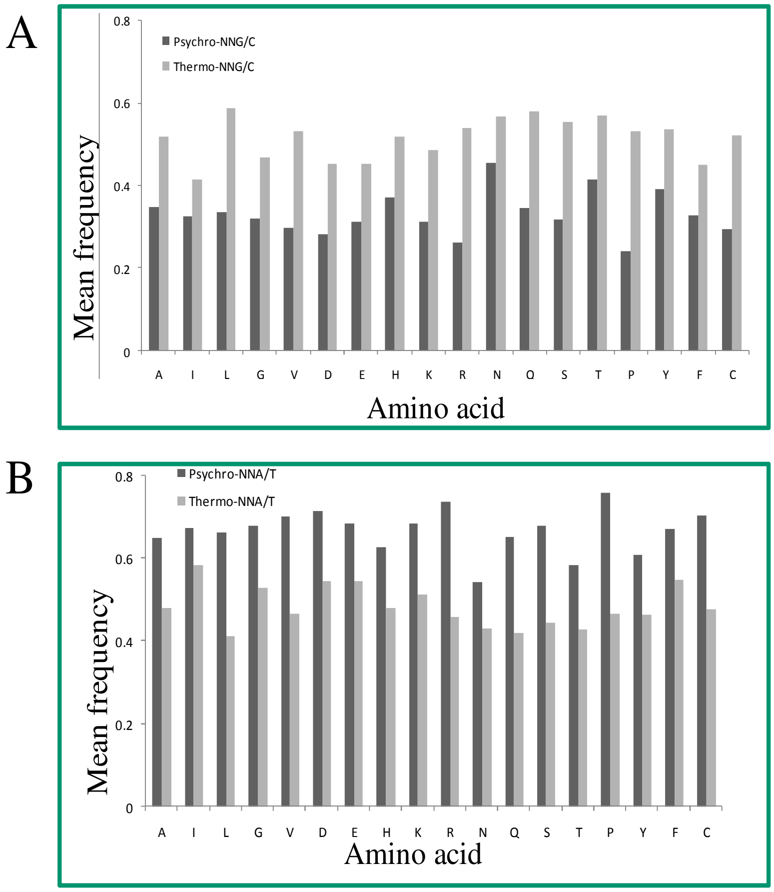


Fig 16: Mean frequency of G/C and A/T ending codons of each amino acid in the conserved genes between thermophiles and psychrophiles. A/T ending codons were seen more in psychrophiles and G/C ending codons were seen more in thermophiles. Amino acids are represented in single letter codes on the *X*-axis and the mean frequency is represented on the *Y*-axis.

Table 9: Normalized frequency of A/T and G/C ending codons of each amino acid of conserved proteins between two extremophiles.

Amino acid	CODONS	Psychrophiles Mean	Thermophiles Mean	t-test(*p*-value)	Amino acid	CODONS	Psychrophiles Mean	Thermophiles Mean	t-test(*p*-value)
ALA	GCT,GCA	0.650	0.480	3.393 (0.0015)	ASN	AAT	0.543	0.431	2.646 (0.0076)
	GCC,GCG	0.350	0.520	-3.393 (0.0015)		AAC	0.457	0.569	-2.646 (0.0198)
ILE	ATT,ATA	0.674	0.585	1.572 (0.0849)	GLN	CAA	0.654	0.42	3.791 (0.0008)
	ATC	0.326	0.415	-1.572 (0.1384)		CAG	0.346	0.58	-3.791 (0.0015)
LEU	TTA, CTT,CTA	0.664	0.412	**4.322 (0.0001)**	HIS	CAT	0.629	0.48	3.474 (0.0008)
	TTG,CTC,CTG	0.336	0.588	**-4.322 (0.0003)**		CAC	0.371	0.52	-3.474 (0.0021)
VAL	GTT,GTA	0.703	0.468	**4.084 (0.0002)**	SER	TCT,TCA,AGT	0.681	0.445	**4.828 (< 0.0001)**
	GTC,GTG	0.297	0.532	**-4.084 (0.0004)**		TCC,TCG,AGC	0.319	0.555	**-4.828 (< 0.0001)**
GLY	GGT,GGA	0.680	0.530	3.488 (0.0006)	THR	ACT,ACA	0.584	0.429	2.708 (0.0078)
	GGC,GGG	0.320	0.470	-3.488 (0.0015)		ACC,ACG	0.416	0.571	-2.708 (0.0155)
ASP	GAT	0.717	0.546	**4.764 (< 0.0001)**	TYR	TAT	0.608	0.463	3.166 (0.0022)
	GAC	0.283	0.454	**-4.764 (< 0.0001)**		TAC	0.392	0.537	-3.166 (0.0054)
GLU	GAA	0.687	0.547	2.840 (0.006)	PHE	TTT	0.672	0.550	2.490 (0.0101)
	GAG	0.313	0.453	-2.840 (0.0118)		TTC	0.328	0.450	-2.490 (0.0207)
LYS	AAA	0.686	0.513	3.049 (0.0038)	CYS	TGT	0.703	0.478	**4.400 (0.0002)**
	AAG	0.314	0.487	3.049 (0.0074)		TGC	0.297	0.522	**-4.400 (0.0002)**
ARG	CGT,CGA,AGA	0.738	0.46	**5.246 (< 0.0001)**	PRO	CCT,CCA	0.759	0.467	**5.244 (< 0.0001)**
	CGC,CGG,AGG	0.262	0.54	**-5.246 (< 0.0001)**		CCC,CCG	0.241	0.533	**-5.244 (< 0.0001)**

Table 9: Normalized frequency of A/T and G/C ending codons of each amino acid of conserved proteins between two extremophiles. A/T ending codons were seen more in psychrophiles and G/C ending codons were seen more in thermophiles. Student t-test shows the significant differences in A/T and G/C ending codons of each amino acid in between thermophiles and psychrophiles. Significant compositional differences are shown in bold between extremophiles.

Preferential selection of GC ending codons by the thermophiles shows that high temperature act as an evolutionary constraint on organisms as G:C pairs are more stable than A:T pairs. Lynn et al. (2002), observed that the optimal growth temperature affects the synonymous codon usage in thermophilic bacteria (Lynn et al. 2002). Mean frequency of G/C ending codons for leucine and valine are more in thermophiles while A/T ending codons are more in psychrophiles. Mean frequency of G/C ending codons for leucine (TTG,CTC,CTG) and valine (GTC,GTG) in thermophiles are 0.588 and 0.532 while in psychrophiles are 0.336 and 0.297 respectively (Table 9). Similarly, mean frequency of A/T ending codons for leucine (TTA,CTT,CTA) and valine (GTT,GTA) in thermophiles are 0.412 and 0.468 while in psychrophiles are 0.664 and 0.703 (Table 9). The other amino acids that are showing highly significant differences ($p<0.0001$) are aspartic acid, arginine, serine, cysteine and proline (Table 9).

Codon usage analyses of groESL also showed similar results with codons ending with A/T were preferred mostly by psychrophiles where as G/C ending codons by thermophiles (Fig 17 and 18). In GroES, cysteine occurrence was seen in two or three thermophilic organisms but not even in a single psychrophile. As frequency of occurrence of that amino acid is too low, it was not considered for t-test calculations (Table 10).

Fig 17: Mean frequency of A/T and G/C ending codons of each amino acid in the GroES gene between thermophiles and psychrophiles.

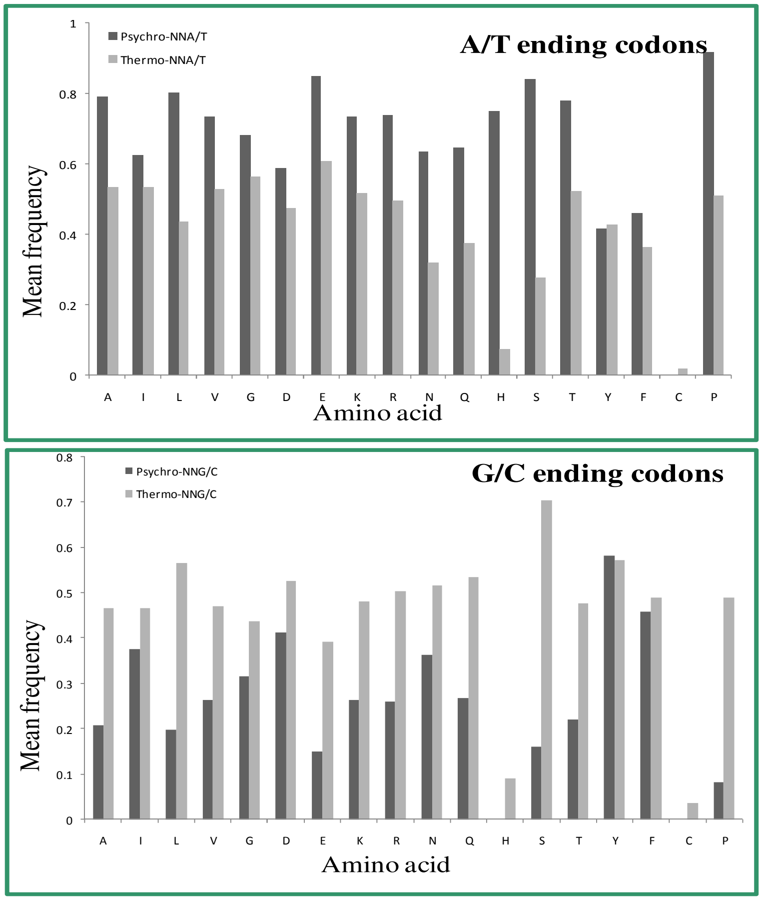


Fig 17: Mean frequency of A/T and G/C ending codons of each amino acid in the GroES gene between thermophiles (grey bar) and psychrophiles (black bar). A/T ending codons were seen more in psychrophiles and G/C ending codons were seen more in thermophiles. Amino acids are represented in single letter codes on the *X*-axis and the mean frequency is represented on the *Y*-axis.

Fig 18: Mean frequency of A/T and G/C ending codons of each amino acid in the GroEL gene between thermophiles and psychrophiles.

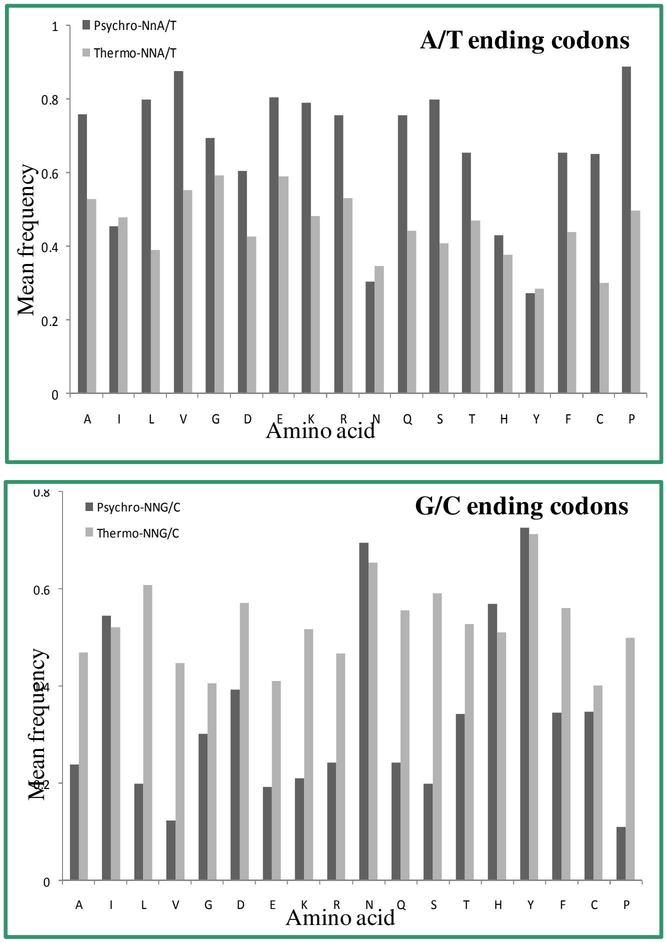


Fig 18: Mean frequency of A/T and G/C ending codons of each amino acid in the GroEL gene thermophiles (grey bar) and psychrophiles (black bar). A/T ending codons were seen more in psychrophiles and G/C ending codons were seen more in thermophiles. Amino acids are represented in single letter codes on the *X*-axis and the mean frequency is represented on the *Y*-axis.

Also, histidine occurrence is very low in either extremophiles. Only A/T ending histidine (CAT) codon usage was seen in psychrophiles with a mean frequency value of 0.750, where as in thermophiles, both A/T and G/C ending histidine codons usage (CAT, CAC) was seen (Fig 17, Table 10). Mean frequency of G/C and A/T ending histidine codon in thermophiles are 0.091 and 0.073 respectively (Table 10). Few amino acids showed highly significant differences ($p<0.00001$) such as alanine, leucine, glutamine, serine and proline with A/T ending codons in psychrophiles and G/C ending codons in thermophiles (Table 10). While in GroEL alanine, luecine, valine, lysine, glutamine, histidine, and proline are showing highly significant differences between the extremophiles (Fig 18, Table 10). Significant differences observed in codon usage are in bold in the Table 9 and 10. The obtained coodn usage showed positive and negative values implies the strength of the selection of the codon usage by that particular group of extremophiles (Table 9 and 10). The negative and/or positive values infer the strength of codon usage bias towards thermophiles and/or psychrophiles respectively which could be due to their adaptive selection of codons. (Table 9 and 10). Lynn et al. (2002), reported that synonymous codon usage was affected by G+C composition of whole genome and the growth at high temperature (Lynn et al. 2002).

Table 10: Normalized frequency of A/T and G/C ending codons of each amino acid of GroESL proteins between two extremophiles.

		Psychro GroES	Thermo GroES		Psychro GroEL	Thermo GroEL	
Amino acid	CODONS	Mean±SD	Mean±SD	t-test(*p*-value)	Mean±SD	Mean±SD	t-test(*p*-value)
ALA	GCT,GCA	0.791±0.153	0.534±0.350	**3.982(0.0002)**	0.760±0.097	0.530±0.348	**4.185(< 0.0001)**
	GCC,GCG	0.209±0.153	0.466±0.350	**-3.982(0.0004)**	0.240±0.097	0.470±0.348	**-4.185(< 0.0001)**
ILE	ATT,ATA	0.624±0.224	0.534±0.277	1.207(0.2015)	0.455±0.245	0.479±0.289	-0.300(0.756)
	ATC	0.376±0.224	0.466±0.277	-1.207(0.2843)	0.545±0.245	0.521±0.289	0.300(0.7663)
LEU	TTA, CTT,CTA	0.802±0.166	0.434±0.313	**5.763(< 0.0001)**	0.800±0.140	0.392±0.329	**6.759(< 0.0001)**
	TTG,CTC,CTG	0.198±0.166	0.566±0.313	**-5.763(< 0.0001)**	0.200±0.140	0.608±0.329	**-6.759(< 0.0001)**
VAL	GTT,GTA	0.735±0.178	0.528±0.311	3.118(0.003)	0.876±0.088	0.553±0.346	**6.039(< 0.0001)**
	GTC,GTG	0.265±0.178	0.472±0.311	-3.118(0.0056)	0.124±0.088	0.447±0.346	**-6.039(< 0.0001)**
GLY	GGT,GGA	0.683±0.147	0.562±0.339	1.931(0.044)	0.697±0.134	0.593±0.291	1.879(0.0631)
	GGC,GGG	0.317±0.147	0.438±0.339	-1.931(0.0763)	0.303±0.134	0.407±0.291	-1.879(0.0606)
ASP	GAT	0.587±0.192	0.472±0.259	1.755(0.0759)	0.606±0.155	0.427±0.233	3.261(0.0042)
	GAC	0.413±0.192	0.528±0.259	-1.755(0.1215)	0.394±0.155	0.573±0.233	-3.261(0.004)
GLU	GAA	0.849±0.113	0.606±0.301	**4.669(< 0.0001)**	0.806±0.167	0.590±0.287	3.475(0.0016)
	GAG	0.151±0.113	0.394±0.301	**-4.669(< 0.0001)**	0.194±0.167	0.410±0.287	-3.475(0.0015)
LYS	AAA	0.734±0.225	0.517±0.294	2.855(0.0077)	0.790±0.180	0.483±0.312	**4.579(< 0.0001)**
	AAG	0.266±0.225	0.483±0.294	-2.855(0.0124)	0.210±0.180	0.517±0.312	**-4.579(< 0.0001)**
ARG	CGT,CGA,AGA	0.740±0.201	0.496±0.359	3.227(0.0024)	0.757±0.145	0.533±0.332	**3.641**(0.0008)
	CGC,CGG,AGG	0.260±0.201	0.504±0.359	-3.227(0.0042)	0.243±0.145	0.467±0.332	**-3.641**(0.0007)

		Psychro GroES	Thermo GroES		Psychro GroEL	Thermo GroEL	
Amino acid	CODONS	Mean±SD	Mean±SD	t-test(*p*-value)	Mean±SD	Mean±SD	t-test(*p*-value)
ASN	AAT	0.636±0.349	0.319±0.417	2.749(0.0118)	0.304±0.172	0.346±0.240	-0.708(0.4329)
	AAC	0.364±0.349	0.517±0.454	-1.300(0.2342)	0.696±0.172	0.654±0.240	0.708(0.4431)
GLN	CAA	0.647±0.421	0.374±0.414	2.042(0.0539)	0.757±0.206	0.443±0.384	**3.965(0.0004)**
	CAG	0.269±0.378	0.535±0.432	-2.147(0.0533)	0.243±0.206	0.557±0.384	**-3.965(0.0004)**
HIS	CAT	0.750±0.452	0.073±0.262	**5.007(0.0003)**	0.800±0.120	0.409±0.310	**7.146(< 0.0001)**
	CAC	-	0.091±0.290	-	0.200±0.120	0.591±0.310	**-7.146(< 0.0001)**
SER	TCT,TCA,AGT	0.840±0.240	0.277±0.360	**6.655(< 0.0001)**	0.656±0.229	0.471±0.376	2.206(0.0375)
	TCC,TCG,AGC	0.160±0.240	0.705±0.371	**-6.376(< 0.0001)**	0.344±0.229	0.529±0.376	-2.206(0.0364)
THR	ACT,ACA	0.779±0.297	0.522±0.390	2.560(0.0154)	0.431±0.423	0.378±0.393	0.396(0.724)
	ACC,ACG	0.221±0.297	0.478±0.390	-2.560(0.0219)	0.569±0.423	0.511±0.412	0.434(0.671)
TYR	TAT	0.417±0.469	0.426±0.376	-0.067(0.9857)	0.274±0.245	0.287±0.253	-0.172(0.7877)
	TAC	0.583±0.469	0.574±0.376	0.067(0.8952)	0.726±0.245	0.713±0.253	0.172(0.798)
PHE	TTT	0.458±0.498	0.364±0.466	0.603(0.5301)	0.655±0.272	0.439±0.285	2.463(0.0268)
	TTC	0.458±0.498	0.491±0.486	-0.206(0.8793)	0.345±0.272	0.561±0.285	-2.463(0.0268)
CYS	TGT	-	0.018±0.134	-	0.653±0.452	0.302±0.413	2.465(0.029)
	TGC	-	0.036±0.188	-	0.347±0.452	0.401±0.448	-0.375(0.6913)
PRO	CCT,CCA	0.917±0.163	0.509±0.339	**6.208(< 0.0001)**	0.890±0.118	0.499±0.387	**6.228(< 0.0001)**
	CCC,CCG	0.083±0.163	0.491±0.339	**-6.208(< 0.0001)**	0.110±0.118	0.501±0.387	**-6.228(< 0.0001)**

Table 10: Normalized frequency of A/T and G/C ending codons of each amino acid of GroESL proteins between two extremophiles. Mean frequency along with their standard deviation (SD) was also mentioned. Student t-test shows the significant differences in A/T and G/C ending codons of each amino acid in between thermophiles and psychrophiles. Significant differences are shown in bold between extremophiles. Dash infers cannot be calculated. As frequency of occurrence of that amino acid is too low therefore it was not considered for t-test calculations.

According to Basak et al. (2004), GC at the third position of codon were found to be significantly higher in thermophillic (*A. aeolicus)* genes than mesophilic *B. subtilis* genes (Basak et al. 2004). It was reported that highly expressed genes prefer to have G or C because of selective pressure at the translation process, while low expressed genes have A or T at the synonymous site (3rd position of codon) since low expressed genes do not suffer such selective pressure as highly expressed genes does (Hou and Yang2003). Variation in the third position of the codon plays an important role to determine the strength of selection of codon usage in case of high temperature loving organisms. High G+C content bias was also observed at the third codon positions in drosophila mitochondrial DNA (Clary and Wolstenholme 1985) and major histocompatability complex class I loci (Hughes and Nei 1988). Usually the codon usage was affected by the various factors including gene expression, G+C composition, strand specific mutational bias, tRNA abundances, transcriptional selection, translational efficiencies (Ermolaeva 2001; Ikemura 1985: Kanaya et al, 1999; Gupta et al. 2004; Ranjan et al. 2007; Sharp et al. 2005). Work done on small number of genes by West and Iglewski 1988, revealed that there is no relation between the codon usage and the gene expression levels (West and Iglewski 1988). Similarly work done in *E. coli* (Ikemura 1981a; Ikemura 1981b; Post and Nomura 1980) and *S. cerevisiae* (Bennetzen and Hall 1982; Ikemura 1982) showed high degree of translational selection depending on codon usage and many have mistakenly assumed that such selection is present in all unicellular organisms. But, recent study has showed that translational selection is not the major factor for the codon usage bias (Lafay et al. 2000). Hou and Yang 2003 showed that the codon usage pattern of high and low expressed genes was partially responsible for the codon usage bias. Irrespective of above reasons, here the differences observed in the codon usage bias across the selected extremophiles could be due to adaptive selection.

4.2.2 To analyze the rate of synonymous and non-synonymous substitutions between thermophilic and psychrophilic conserved orthologs to understand the molecular evolution of either extremophiles:

We investigated the nature of nucleotide substitutions, which might have occurred and accepted during the course of adaptation to either temperature extremes. SNAP calculates pair wise synonymous and non-synonymous distances according to the Nei and Gojobori method for an alignment in table format (Nei and Gojobori 1986). It outputs a synonymous and a non-synonymous distance matrix, an exhaustive pairwise codon-by-codon comparison between the pair of aligned sequences. The ratio of non-synonymous substitutions versus the synonymous substitutions, d_N/d_S is widely used to estimate the evolutionary relationship and to measure the selection pressure acting on protein coding region. The available methods can be used to calculate the non-synonymous and synonymous substitution rate d_N/d_S of genes. We introduced a new method for calculating the conserved non-synonymous (CNS) and non-conserved non-synonymous substitutions (NCNS) including synonymous substitutions. Conserved non-synonymous substitution (CNS) is substitution of an amino acid with another amino acid having similar physicochemical properties. Such a conserved substitution has a greater chance of being accepted during the course of evolution. While non-conserved non-synonymous substitution (NCNS) is substitution of an amino acid with another amino acid having different physicochemical proteins. The calculation for identifying the substitution rate by our designed approach was mentioned clearly in methodology. d_{NCNS}/d_{SYN} ratio was calculated by estimating the number of non-conserved non-synonymous substitution per non-conserved non-synonymous site (d_{NCNS}) and the number of synonymous substitution per synonymous site (d_{SYN}). Similarly, d_{CNS}/d_{SYN} was calculated by estimating the number of conserved non-synonymous substitution per conserved non-synonymous site (d_{CNS}) and the number of synonymous substitution per synonymous site (d_{SYN}).

The 3D plot of mean of synonymous, conserved and non-conserved non-synonymous substitutions obtained by considering the 55 thermophiles and 12 psychrophiles of conserved proteins was shown in

Fig 19. Compared to conserved non-synonymous substitution, non-conserved non-synonymous substitution was seen more within and between the selected extremophiles. Whereas, conserved non-synonymous substitution was almost same between the extremophiles (Fig 19). The average d_{SYN} and d_{NCNS} in psychrophiles are 0.650 ± 0.158; 0.217 ± 0.05 where as in thermophiles are 0.462 ± 0.086; 0.294 ± 0.034 respectively. These differences in the d_{SYN} and d_{NCNS} between extremophiles infer that organisms undergo certain substitutions might be necessary for their adaptation during the course of evolution. High temperature imposes more constraints on the genome of thermophiles than psychrophiles. Therefore, the non-conserved non-synonymous substitutions were seen more in thermophiles than psychrophiles to cope up with high temperature selection pressure though these are not acceptable, but might be necessary for their adaptation to high temperature. The number of d_{CNS} were almost same between thermophilic and psychrophilic proteins inferring that during course of evolution conservative substitutions have a greater chance of being acceptable (Fig 19) (French and Robson 1983). Synonymous substitution rate study in thermophilic and mesophilic bacteria showed a decrease in d_S for the information processing genes for *M. jannaschii* and *M. maripaludis* suggested that synonymous codon usage was subjected to greater constraint than other functional genes (Basak and Ghosh 2006).

Fig 19: 3D Plot of mean of synonymous substitutions (d_{SYN}), conserved (d_{CNS}) and non-conserved (d_{NCNS}) non-synonymous substitutions observed between thermophilic and psychrophilic conserved proteins.

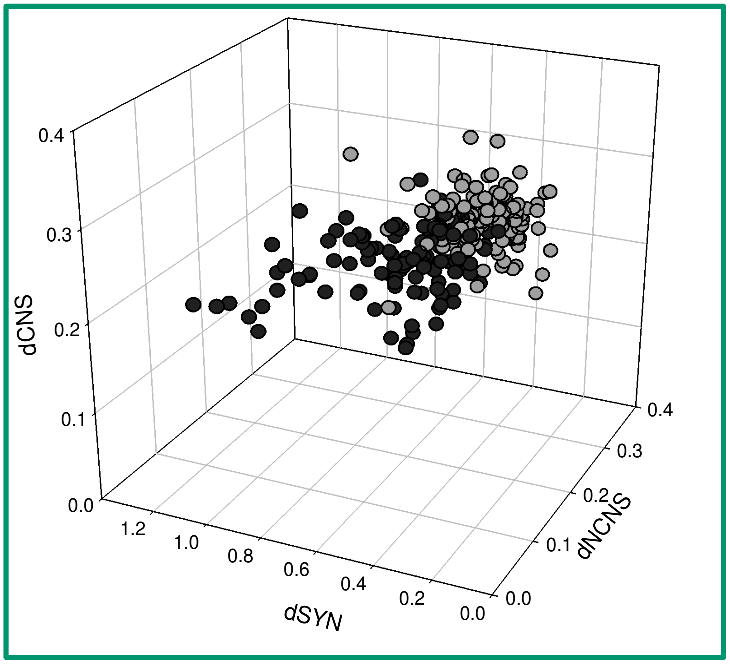


Fig 19: 3D Plot of mean of synonymous substitutions (d_{SYN}), conserved (d_{CNS}) and non-conserved (d_{NCNS}) non-synonymous substitutions observed between thermophilic and psychrophilic conserved proteins. Non-conserved non-synonymous substitutions more in thermophiles (grey cirlce) and synonymous substitution more in psychrophiles (black circle). While conserved substitution were same between both extremophiles. X-axis represents synonymous substitution (d_{SYN}); *Y*-axis represents conserved non-synonymous substitution (d_{CNS}) and *Z*-axis represents non-conserved non-synonymous substitution (d_{NCNS}).

Recently, according to Yu et al. (2012), average d_N/d_S values of conserved genes in *Synechococcus* are significantly higher than *Prochlorococcus*. Synonymous codon usage bias was also analyzed in the cyanobacterial species *Synechococcus* and *Prochlorococcus* and found that the differences observed could be due to genome compositional constraints or their adaptation to the environments (Yu et al. 2012). If the rate of non-synonymous substitution is similar to the rate of synonymous substitution it means that substitution results in no effect on the function of the protein. The average d_{NCNS}/d_{SYN} of conserved proteins of thermophiles (0.686 ± 0.408) is significantly higher than psychrophiles (0.476 ± 0.296) (Fig 19). Higher substitutions in thermophiles might be beneficial for the species to survive in this high temperature niches.

Similar results were observed in GroES and GroEL between the two extremophiles. Non-conserved non-synonymous substitutions were more in thermophilic GroESL and synonymous substitution were more in psychrophilic GroESL (Fig 20). The average d_{NCNS}/d_{SYN} of GroES and GroEL proteins of thermophiles (0.947 ± 0.515; 0.515 ± 0.167) is significantly higher than psychrophiles (0.263 ± 0.076; 0.165 ± 0.086) (Fig 20).

Fig 20: 3D Plot of mean of synonymous substitutions (d_{SYN}), conserved (d_{CNS}) and non-conserved (d_{NCNS}) non-synonymous substitutions observed between thermophilic and psychrophilic GroESL proteins.

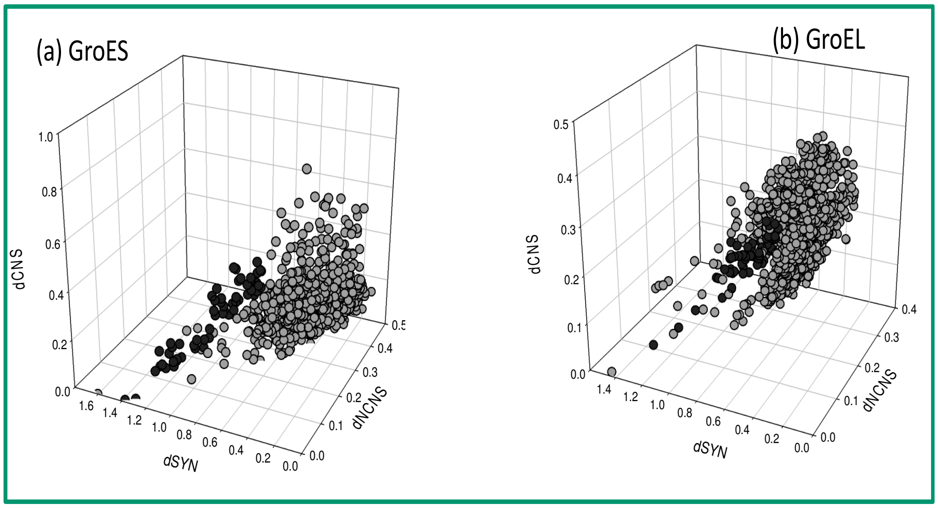


Fig 20: 3D Plot of mean of synonymous substitutions (d_{SYN}), conserved (d_{CNS}) and non-conserved (d_{NCNS}) non-synonymous substitutions observed between thermophilic and psychrophilic GroESL proteins. 3D plot shows non-conserved non-synonymous substitutions more in thermophiles (grey circle) and synonymous substitution more in psychrophiles (black circle). While conserved substitution were same between both extremophiles. X-axis represents synonymous substitution (d_{SYN}); *Y*-axis represents conserved non-synonymous substitution (d_{CNS}) and *Z*-axis represents non-conserved non-synonymous substitution (d_{NCNS}).

Identifying the non-synonymous and synonymous sites in the large data set of selected extremophiles is difficult as most of the amino acids are conserved and positions that are altering as adaptive selection are difficult to correlate between thermophiles and psychrophiles. Large data set was considered for identification of genes on which positive selection operate by Endo et al. 1996. They identified only 17 proteins out of 3,595 as likely to be under positive selection, at a proportion of only 0.47 % (Endo et al. 1996).

Substitution rate analysis can, therefore, help us to identify whether the organisms are under positive or negative selection. This result clearly demonstrates that thermophiles have undergone a higher rate of substitutions which is necessary for their thermal adaptability. In other words, non-synonymous substitutions took place at a faster rate than synonymous substitution in thermophiles. This indicates that thermophiles have undergone positive selection while psychrophiles have underogone negative selection. Positive selection is a substitution which occurs with the increased frequency of preferred codons where as negative or purifying selection is the substitution with decreased frequency of preferred codons. Rate of evolution depends on the degree of selective constraints, stronger the constraint slower the rate of molecular evolution. Thereby, it might be that cold loving organisms are under stronger constraints and thus showed slower rate of evolution compared to thermophiles, which might have undergone higher substitution rates. It can infer that thermophiles have strong compositional constraints, which have undergone much more substitution than psychrophiles and these substitutions are necessary for their adaptation to extreme temperature.

4.3 Conclusion:

Analysis of 119 conserved genes showed considerable difference in GC content, codon usage and synonymous and non-synonymous substitution rate between the thermophiles and psychrophiles. As thermophiles have been evolved to survive in extremely high temperatures, it is reasonable to have high GC content than psychrophiles. Codon usage also showed variation between the extremophiles with A/T ending codons in psychrophiles and G/C ending codons in thermophiles. Stronger selection of GC ending codons by the thermophiles shows that high temperature act as evolutionary constraint on organisms. Thermophiles experience more constraints on the genome than psychrophiles. As a result they might have undergone non-conserved substitutions, which are not naturally acceptable, but, might have been carried forward for adaptation to high temperature. While conserved substitution was equally distributed between extremophiles as during course of evolution conservative substitutions have a greater chance of being acceptable. These results prove that thermophiles have undergone positive selection and psychrophiles have undergone negative selection.

Chapter 5:

In order to see whether mechanism of gene regulation is optimized to sustain either extremes of temperature, the *groE* operon regulation was analyzed in this present chapter. Nature of molecular changes that were observed in the proteins and the genes of *groESL* have further led to check for the differences in the *groE* operon regulation between thermophiles and psychrophiles. In bacteria genes are regulated by the *cis* regulatory element located mostly in the upstream region of gene. We identified different *cis* regulatory elements in the upstream of *groES* gene between thermophiles and psychrophiles. Transcriptional regulators of *groE* operon were compared between psychrophiles and thermophiles to see any differences in the regulatory mechanism to sustain either extremes temperature. GroESL, one of the heat shock protein was well studied and characterized in most of the bacteria are regulated by the inverted repeat (CIRCE element) or by the alternative sigma factor.

5.1 Introduction:

5.1.1 Heat shock protein regulation in Prokaryotes:

In *Bacillus subtilis,* three classes of heat shock regulons have been defined which are regulated by different mechanisms (Hecker et al. 1996). Class I heat shock genes are regulated by the inverted repeat called CIRCE element which serve as the binding site for the repressor protein and by the vegetative promoter P_A (Zuber and Schumann 1994; Chang et al.

1994). The genes that comes under the class I includes *dnaK* and *groE* operons. In *clostridium acetobutylicum, staphylococcus aureus, lactobcoccus lactis* and *chlasmydia trachomatis* CIRCE sequence was reported in upstream of *dnaK* operon (Eaton et al. 1993; Narberhaus et al. 1992; Ohta et al. 1994). Class II heat shock genes are positively regulated by the alternative σ factor. Some heat shock genes, which neither belongs to class I nor class II, have been grouped into class III. Upon heat shock, 66 heat shock proteins are detected in *B. subtilis*. Among which the functional role of GroES/GroES, DnaK, DnaJ and GrpE facilitate protein folding, assembly and export and minimizes the protein aggregation (Gragerov et al. 1992; Horwich et al. 1993; Schröder et al. 1993).

In *E. coli,* DnaK is regulated by chaperone protein DnaJ and by cofactor GrpE. *grpE*, *dnaK* and *dnaJ* exist in an operon which is constitutively required for its growth at all temperatures (Roberts et al. 1996). This operon functions in a multi protein complex as a molecular chaperon to assist in proper protein folding. DnaK was assumed to be involved in prevention and repair of damaged proteins. The co-chaperone GrpE (for GroP-like E) functions as nucleotide exchange factor which releases adenosine 5` phosphate (ADP) for ATP from DnaK allowing the folded protein to be released (Szabo et al. 1994). GrpE present in mitochondria, chloroplast and cytoplasm which act as sensor molecules and was found to be expressed in all temperature conditions irrespective of heat shock. Several GrpE mutants have been generated to understand the role in protein folding, its interactions with co-chaperones and the functional residues necessary for its efficient activity (Harrison et al. 1997).

The *hrcA*, first gene of the *dnaK* operon in *B. subtilus* encodes a repressor protein that controls the expression of *dnaK* and *groE* operons (Roberts et al. 1996; Schulz and Schumann 1996). In *Bacillus dnaK* operon consists of four genes in the order of *hrcA*-*grpE*-*dnaK*-*dnaJ* (Wetzstein et al. 1992). In most of the microorganisms the *dnaK* operon was not completely sequenced. HrcA act as repressor which binds specifically to CIRCE element or as transcriptional activator for its own gene expression. In *E. coli* neither HrcA nor CIRCE element was reported in regulation of heat shock genes. Disruption of entire *dnaK* operon in *B. subtilis* resulted in higher expression of *groE*, while the inactivation of either *dnaK* or *DnaJ* or both does not affect the *groE* expression (Schulz et al. 1995). Mutations in one or both the arms of this inverted repeat of the CIRCE element resulted in elevated transcription of *groESL* and *dnaK/dnaJ/grpE* genes even at normal growth temperature (Zuber and Schumann 1994).

5.1.2 Role of *cis* regulatory elements:

Understanding the mechanism of regulation of genes, DNA binding elements and its regulatory sites at the molecular level are the greatest challenges in modern biology. Genes are up-regulated when it is required and down-regulated when it is not required, for example the Lac operon genes are up-regulated when the lactose is available in the environment and switched off, when lactose is not available in the environment (Oehler et al. 1990). The most common mechanism to understand gene expression operates at the stage of transcription. RNA polymerase complex first binds to the promoter region to produce mRNA transcript from a gene template. Transcription factors bind to the promoter region that can either activate or repress by promoting or interfering with transcription process. Regions of DNA or RNA that regulate the expression of genes located on that same strand are called *cis*-regulatory elements which often forms the binding sites of one or more trans-acting factors. *Cis* regulatory elements are difficult to identify due to great variation in the binding site (Spivakov et al. 2012). It was known that, conservation of orthologous genes or co regulated genes within the closely related organism, have conservation in non coding upstream sequence which form the binding sites of c*is* regulatory element (McGuire et al. 2000). Thus these non coding regions containing functionally significant regions are necessary for the regulation of genes and have been retained with the gene. Organisms with shared orthologs have similar function therefore, share similar mechanism of gene regulation. Conserved transcription factor binding sites reported in humans (Whitfield et al. 2012), bacteria (McGuire et al. 2000) and yeast (Pritsker et al. 2004) support the augment that important residue or regions necessary for regulation are retained during the evolution of transcription factor binding site. Similar approach of identification of *cis* regulatory elements by comparative genomics between *Lactobacillus* and *Bacillus* species was done by Wels et al. (2006), where they identified the common, *Bacillus* and *Lactobacillus* specific regulatory motifs (Wels et al. 2006). Similarly, phylogenetic footprinting by Tan et al. (2001), have been applied in comparison between *E. coli* and *H. influenzae* genomes where they identified a novel motif that had not been found previously in any of the individual genomes (Tan et al. 2001). Liu et al. (2008), presented a new approach for genome wide identification of c*is*-regulatory motifs in bacterial systems and uncovered the c*is* regulatory map of *Shewanella oneidensis* (Liu et al. 2008).

5.2 Results and Discussion:

5.2.1 *Cis* regulatory element in the upstream of *groE* operon:

Upstream sequences are retrieved from Uniprot (http://www.Uniprot.org/) for the orthologous genes of *groES* of selected extremophiles for identification of c*is* regulatory elements. The identified motifs were validated by the published literature on the c*is* elements.

5.2.1.1 In Thermo *groE* operon:

The conserved *cis* regulatory elements were identified by submitting the upstream sequences of *groESL* to MEME (Multiple Expectation maximum for Motif Elicitation). CIRCE element was identified in 48 out of 55 thermophilic *groES* upstream (Fig 21). In few thermophiles, two CIRCE elements were identified which include *A. cellulolyticus* 11B, *C. aurantiacus* J-10-fl, *C. aggregans* DSM 9485 and *R. castenholzii* DSM 13941 (Fig 22). However, CIRCE element was not identified in the upstream sequence of any of psychrophilic *groES*.

CIRCE element is a nanomeric inverted repeat of 9 bp separated by 9 bp spacer which is highly conserved operating sequence in the bacterial species (Zuber and Schumann 1994). CIRCE (for Controlling the Inverted Repeat of Chaperone Expression) element TTAGCACTC-N9-GAGTGCTAA is usually seen in the upstream of *groESL* operon, *dnaK* operon of several gram negative bacteria. Identified CIRCE element has a consensus sequence of 5’-TTAGCACTC-(N9)-GAGTGCTAA-3’ in thermophiles which exactly matched with the reported CIRCE element (Zuber and Schumann 1994). Few thermophiles such as *S. azorense* Az Fu1, *D. geothermalis* DSM 11300, *A. aeolicus* VF5, *Hydrogenobaculum* sp. YO4AAS1, *Sulfurihydrogenibium* sp. YO3AOP1, *C. proteolyticus* DSM 5265 and *M. infernorum* V4 showed no CIRCE element. Presence of CIRCE element only in thermophiles and its absence in psychrophiles infers that thermophiles differ in the regulation of *groESL* gene from psychrophiles.

Fig 21: Identified inverted repeat in the upstream of *groES* orthologs from thermophiles.

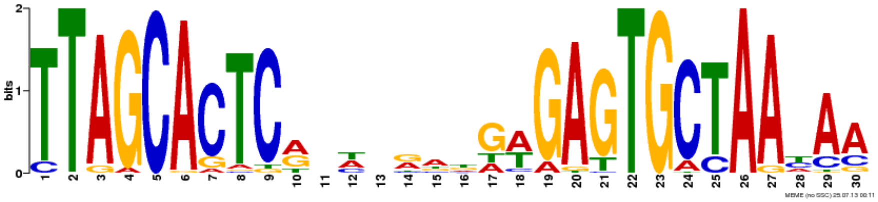


Fig 21: Identified inverted repeat in the upstream of *groES* orthologs from thermophiles. Consensus of inverted repeat (CIRCE element) predicted by MEME upon submission of 55 thermophilic *groES* upstreams. The default parameters used for motif prediction include repeat of minimum length of 6 bp and maximum of 30 bp nucleotide and loop region of less than 10 nucleotide length. This motif is known to be the binding site of HrcA in both gram negative and gram positive bacteria. This element is highly conserved in thermophilic *groES*, but is absent in the upstreams of psychrophilic *groES.*

Fig 29: Schematic representation of CIRCE element and promoter -10 and -35 sites in the upstream sequence of *groES* orthologs of thermophiles.

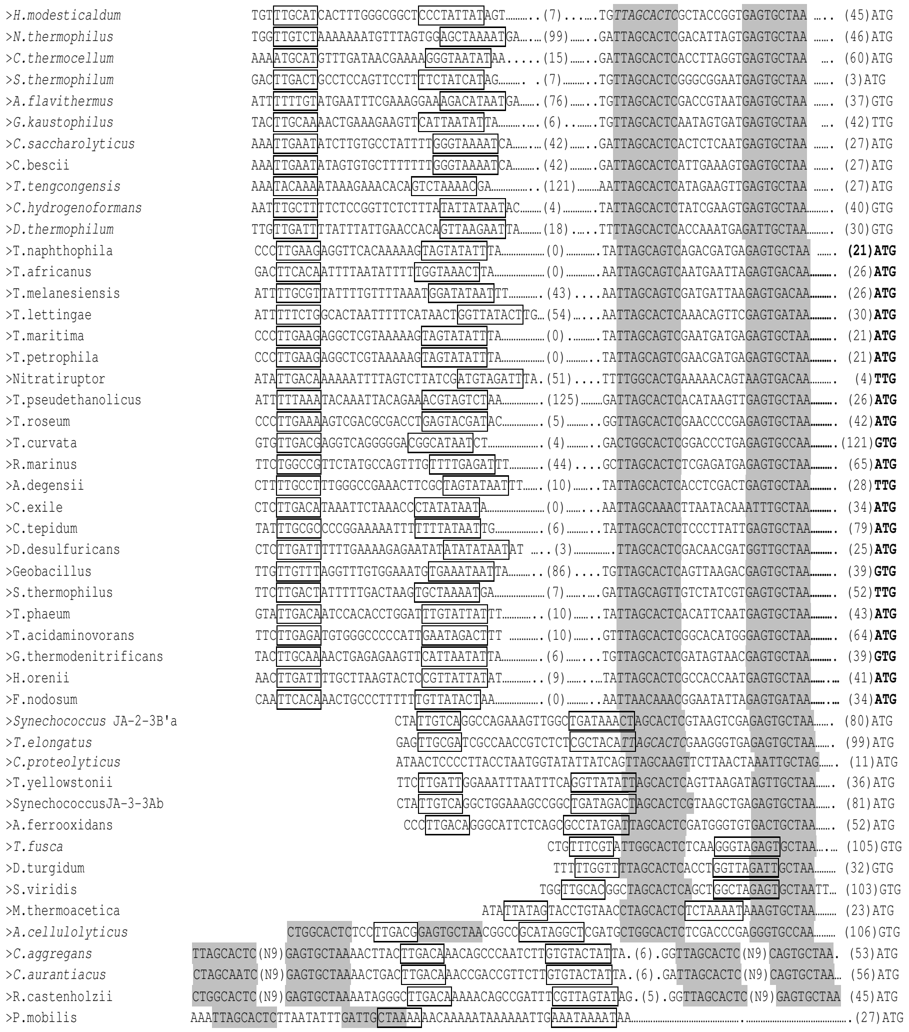


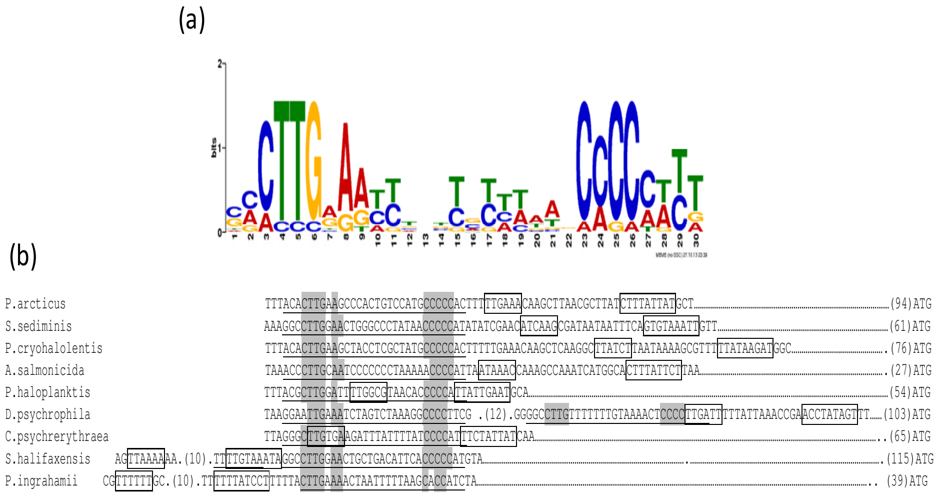
Fig 29: Schematic representation of CIRCE element and promoter -10 and -35 sites in the upstream sequence of *groES* orthologs of thermophiles. The positions of CIRCE element and promoter sites from ATG in the upstream of thermophilic *groES* were shown here. Inverted repeat of CIRCE element with right and left nine bases are highlighted in grey and -10 and -35 promoter site are shown in boxes. CIRCE element was identified by MEME and promoter by BPROM online software.

Regulation of *groESL* gene differs between gram positive and gram negative bacteria. In gram positive bacteria *B. subtilus* and *C. acetabutylicum* a highly conserved inverted repeat known as CIRCE element controls the regulation of *groE* and *dnaK* operon (Hecker et al. 1996; Narberhaus and Bahl 1992). In gram negative bacterium *E. coli* the heat shock genes are positively regulated by an alternative sigma factor σ^{32} and negatively regulated by the products of heat shock genes, *dnaK*, *dnaJ* and *grpE* (Yura et al. 1993). Although thermophiles include both gram positive and negative bacteria CIRCE element was identified in the upstreams of *groES* ORFs irrespective of whether they belong to gram positive or gram negative bacteria.

5.2.1.2 In Psychro *groE* operon:

Psychrophiles considered here were all gram negative organisms. However few reports say that, in several gram negative bacteria, CIRCE element was also found in the upstream of *groE* operon along with sigma 32 like promoter site (Schulz and Schumann 1996). Instead of CIRCE element, RpoH binding site was predicted in front of *groES* sequence of 9 psychrophiles by MEME (Fig 23 a and b). In psychrophiles, *groESL* gene regulation might be by sigma 32 encoded by *rpoH* gene.

Fig 23: Motif identified in the upstream of *groES* genes of psychrophiles.



```
P.arcticus                   TTTACACTTGAAGCCCACTGTCCATGCCCCCACTTTTTGAAACAAGCTTAACGCTTATCTTTATTATGCT.......(94)ATG
S.sediminis                  AAAGGCCTTGGAACTGGGCCCTATAACCCCCATATATCGAACATCAAGCGATAATAATTTCAGTGTAAATTGTT....(61)ATG
P.cryohalolentis             TTTACACTTGAAGCTACCTCGCTATGCCCCCACTTTTTGAAACAAGCTCAAGGCTTATCTTAATAAAAGCGTTTTTATAAGATGGC.(76)ATG
A.salmonicida                TAAACCCTTGCAATCCCCCCCTAAAAACCCCATTAATAAACCAAAGCCAAATCATGGCACTTTATTCTTAA......(27)ATG
P.haloplanktis               TTTACGCTTGGATTTTGGCGTAACACCCCCATTATTGAATGCA..................................(54)ATG
D.psychrophila               TAAGGAATTGAAATCTAGTCTAAAGGCCCCTTCG .(12).GGGGCCTTGTTTTTTTGTAAAACTCCCCTTGATTTTTATTAAACCGAACCTATAGTTT.....(103)ATG
C.psychrerythraea            TTAGGGCTTGTGAAGATTTATTTTATCCCCATTTCTATTATCAA.................................(65)ATG
S.halifaxensis    AGTTAAAAAA.(10).TTTTGTAAATAGGCCTTGGAACTGCTGACATTCACCCCCATGTA..........(115)ATG
P.ingrahamii     CGTTTTTTGC.(10).TTTTTTATCCTTTTTACTTGAAAACTAATTTTTAAGCACCATCTA.......... (39)ATG
```

Fig 23: Motif identified in the upstream of *groES* genes of psychrophiles. (a) Consensus sequence of RpoH binding site predicted by MEME upon submission of 12 psychrophilic *groES* upstreams. (b) Figure shows the positions of RpoH binding site and promoter sites from ATG. RpoH binding site with right and left four bases are highlighted in grey and -10 and -35 promoter site are highlighted in boxes. This element is highly conserved in psychrophilic *groES*, but is absent in the upstreams of thermophilic *groES*. RpoH binding site was identified by MEME and promoter by BPROM online software.

The identified consensus sequence of the RpoH binding site was 5"-CTTGAAAC(N_{12})CCCCATAA3' which is similar to RpoH binding site identified in *E. coli* (5'-ggcCTTG(N_{12-20})CCCCAT-3') and *N.*

gonorrhoeae (5’-CTTGaaatt($N_{9\text{-}12}$)CCNNatttt-3’) (Gunesekere et al. 2006). Conserved regions was underlined and found to be located at -35 and -10 promoter site identified by BPROM program (Fig 23b). RpoH gene encodes sigma 32 factor which recognizes the promoter for genes responding to heat shock. Sigma 32 promoter site was also identified by the primer extension analysis in psychrophilic bacterium C. *psychrerythraea* in the upstream of *groE* operon (Yamauchi et al. 2012)

In *E. coli,* low levels of *rpoH* are necessary for the transcription of chaperones involved in refolding of newly synthesized proteins but, increase drastically (about 20-fold) during the first few minutes after heat shock (Rosen et al. 2002; Yura et al. 1993). The regulation of *rpoH* is controlled at posttranscriptional level, where translation of the protein is repressed by secondary structure in the *rpoH* mRNA (Yuzawa et al. 1993; Laskos et al. 2004). The level and activity of σ32 are negatively regulated by the products of the heat shock genes *dnaK*, *dnaJ*, and *grpE* (Herman et al. 1995; Tomoyasu et al. 1995*).* DnaK and DnaJ inhibit σ32 activity by preventing its binding to the RNA polymerase core while GrpE partially reverses this inhibition (Gamer et al. 1996). In another example *Caulobacter crescentus*, a gram-negative bacterium several heat shock-inducible genes are found to have σ32 like promoters (Avedissian and Gomes 1996; Gomes et al. 1990).

When we checked for the presence of *ropH* gene orthologs in psychrophiles and thermophiles, orthologs for *rpoH* gene was seen only in 12 psychrophiles and not even in a single thermophile. Thus, our data clearly suggests that in psychrophiles, *groESL* operon is regulated by sigma 32 dependent activation mechanism, while in thermophiles, it is by CIRCE/HrcA system.

5.2.2 *DnaK* operon architecture in selected extremophiles:

CIRCE element forms the binding site for the repressor protein *HrcA* (heat regulation at CIRCE) encoded by orf39, the first gene of *dnaK* operon (Jäger et al. 2004). HrcA protein binds to the CIRCE element at DNA level and serves as repressor for *groE* and *dnaK* operon in *B. subtilus* (Schulz and Schumann 1996*).* We have checked for the presence of *HrcA* gene in the selected extremophiles, HrcA orthologs were seen in majority of thermophiles but, not in psychrophiles except *P. haloplanktis* (Fig 24). Few thermophiles also showed no *hrcA* gene which includes *S. azorense* Az Fu1, *D. geothermalis* DSM 11300, *A. aeolicus* VF5, *Hydrogenobaculum* sp. YO4AAS1, *Sulfurihydrogenibium* sp. YO3AOP1, *C. proteolyticus* DSM 5265 and *M. infernorum* V4. These seven organisms also showed no CIRCE element in front of *groESL* gene. To assess the extent of importance of this analysis we further checked for the presence of CIRCE element in the whole genome sequence of all psychrophiles and thermophiles by MAST (http://meme.nbcr.net/meme/cgi-bin/mast.cgi). It was also noticed that CIRCE element was observed in the non coding part of *hrcA* gene of all thermophiles only. Apart from it, CIRCE element was also identified in front of other small heat shock genes like Hsp 90, Hsp 100 etc., in thermophiles, but not in psychrophiles. Where as in psychrophiles, no structural CIRCE element was identified in the whole genome sequence of heat shock genes, which might be another reason for the differences in the gene regulation of heat shock genes between psychrophiles and thermophiles.

We looked into *dnaK* operon architecture between thermophiles and psychrophiles. Usually *dnaK* operon consists of four genes in the order of *HrcA-GrpE-DnaK-DnaJ* (Wetzstein et al. 1992) in *B. subtilus* and *S. coelicolor* (Bucca et al. 1997) and in most of the other bacteria. Later in 1997, *dnaK* operon was found to be heptacistronic in *B. subtilus*, with three new genes (*orf35*, *orf28* and *orf50*) found to be transcribed along with *dnaK* operon genes (Homuth et al. 1997). In both thermophiles and psychrophiles multiple copies of *dnaK*, *dnaJ* and *grpE* genes were seen in the genomes. In thermophiles four gene operon *hrcA-grpE-dnaK-dnaJ* was seen in most of the cases where as in psychrophiles, only the two gene operon *dnaK-dnaJ* or *grpE-dnaK* was mostly observed with poor gene order and variation in the operon architecture or synteny (Fig 24).

Fig 24: Schematic representation of *dnaK* operon architecture in thermophiles and psychrophiles.

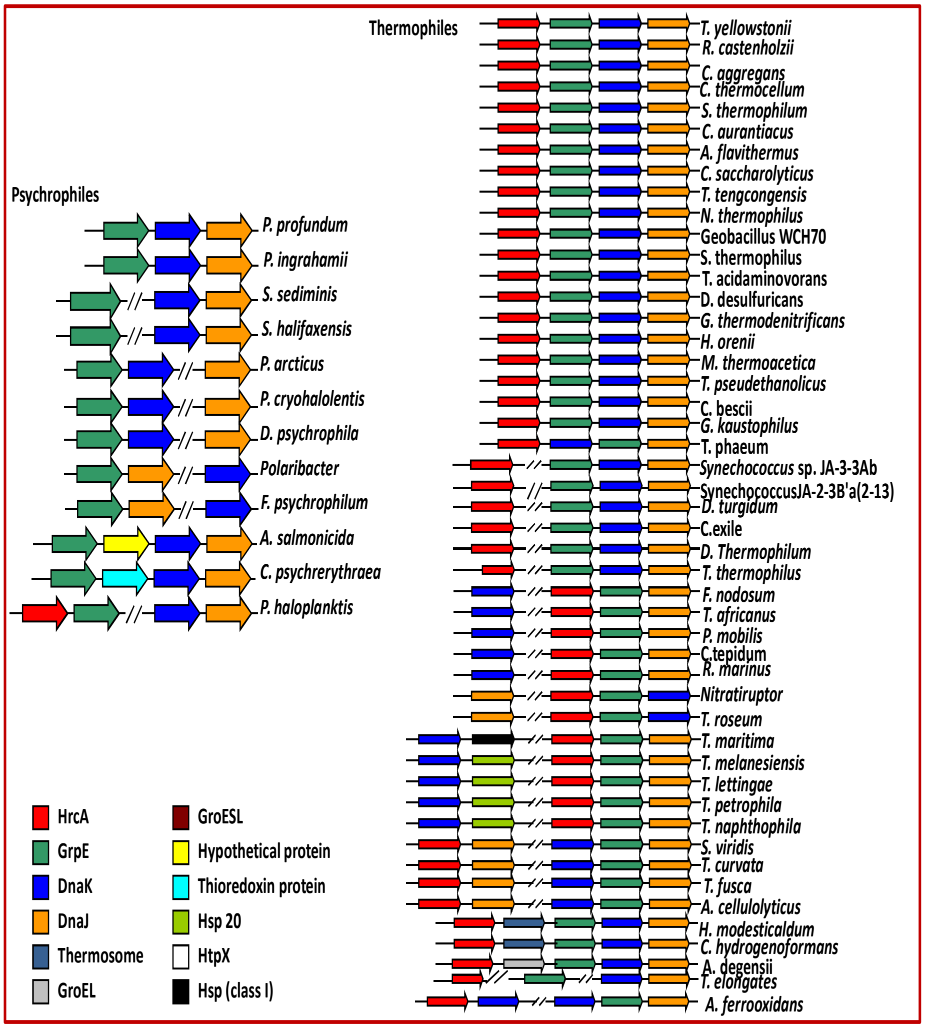


Fig 24: Schematic representation of *dnaK* operon architecture in thermophiles and psychrophiles. *dnaK* operon architecture with four genes (*hrcA*-*grpE*-*dnaK*-*dnaJ*) in thermophiles and two genes (*grpE*-*dnaK or dnaK*-*dnaJ*) in psychrophiles. *HrcA* gene was not seen in psychrophiles where as in few cases, *hrcA* gene in not linked with the *dnaK* operon in thermophiles.

In few cases the *GrpE* gene was found to be not linked to the *dnaK* operon in psychrophiles and this forms a good example for the gene rearrangement due to the diversity of extremophiles (Fig 24). In *H. modesticaldum, C. hydrogenoformans, A. degensii* five genes *dnaK* operon was identified with a transposone gene found next to *HrcA* in *H. modesticaldum* and *C. hydrogenoformans* where as in *A. degensii groEL* gene was found next to *hrcA* gene. In few thermophiles, *dnaK* and *hrcA* was found not to be linked to the *dnaK* operon. Neither *hrcA* gene nor CIRCE element was seen anywhere in the entire genome of all psychrophiles.

5.2.3 *Cis* regulatory element in the upstream of *dnaK* operon:

HrcA first gene in *dnaK* operon in thermophiles and *GrpE,* the first gene of *dnaK* operon in psychrophiles were considered for *cis* element predictions. Upstream sequence of GrpE and *HrcA* orthologs of psychrophiles and thermophiles were retrieved from the Uniprot and submitted to MEME for the identification of motif. It was observed that CIRCE binding site was identified in the noncoding region of HrcA orthologs in thermophiles where as RpoH binding site was identified in the noncoding region of GrpE orthologs in psychrophiles. CIRCE element was identified in 39 thermophiles out of 47 which are having HrcA gene. As upstream intergenic sequence was too small for the *T. thermophilus* it was not considered for *cis* regulatory element identification The consensus sequence of the RpoH binding site was 5”-CTTGAAAC(N_{12})CCCCATAA3’ which is similar to the RpoH binding site identified in *E. coli* and *N. gonorrhoeae* (Gunesekere et al. 2006). Conserved region was underlined and was found to be located near the -35 and -10 position identified by BPROM program. It might be that *GrpE* gene is under the regulation of sigma 32 factor in case of psychrophiles.

5.2.4 Evolutionary basis of adaptations:

Absence of this CIRCE element in psychrophiles and RpoH binding site in thermophiles infers that during the course of adaptation, resulted substitution or deletion in noncoding region and retained only those regions that are necessary or suitable for their adaptation. It can be assumed that as they experience high temperature, expression and accumulation of heat shock proteins might go beyond the required amount in the cell. Therefore, they are evolved with mechanism of repression to control the HSP levels in the cell, whereas psychrophiles, can activate the expression, as and when they experience high temperature, by *rpoH* dependent activation, as they rarely experience high temperatures in their habitat. These results prove that thermophiles have undergone positive selection and psychrophiles have undergone negative selection. Maintaining the positive selection in psychrophiles and negative selection in thermophiles of *groESL* gene regulation infers that effective selection was preferred during their adaptations to extreme conditions.

5.3 Conclusion:

We have observed that the mechanism of *groESL* gene regulation is different between thermophiles and psychrophiles. In most of the thermophiles, majority of heat shock genes and operons are found to be regulated by CIRCE / HrcA system. However, in the genome of psychrophiles, CIRCE element and *hrcA* gene were not found. Instead, RpoH binding site was found in the upstreams of *groE* and *dnaK* operons in psychrophiles. Absence of CIRCE element and *hrcA* gene in psychrophiles infers that negative regulation of heat shock genes might not be necessary as they are rarely exposed to heat. They probably need an immediate activation of *hsp* gene expression upon heat shock. However, thermophiles experience high temperatures all the time, to maintain the mRNA and protein levels of heat shock proteins in the cell they might have evolved with negative regulatory mechanism. Overall, the data generated, positive regulation in psychrophiles and negative regulation in thermophiles infers that they have selected suitable mechanism for the regulation of genes during adaptation to extreme temperature conditions.

References:

Abbott JC, Aanensen DM, Bentley SD, 2007. WebACT: an online genome comparison suite Methods. *Mol Biol*. 395: 57-74.

Aguilar CF, Sanderson I, Moracci M, Ciaramella M, Nucci R, Rossi M, Pearl LH. 1997. Crystal structure of the β-glycosidase from the hyperthermophilic Archaeon *Sulfolobus solfataricus*: resilience as a key factor in themostability. *J Mol Biol.* 271: 789-802.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics*. 136: 927-35.

Akashi H. 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. *Gene.* 238: 39-51.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215: 403–410.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. *Microbiol Rev.* 54: 198-210.

Avedissian M, Gomes SL. 1996. Expression of the *groESL* operon is cell cycle controlled in *Caulobacter crescentus. Mol Microbiol*. 19: 79–89.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*. 2: 28- 36.

Baker-Austin C, Dopson M. 2007. Life in acid: pH homeostasis in acidophiles. *Trends Microbiol.* 15: 165-71.

Bakermans C, Skidmore ML. 2011. Microbial metabolism in ice and brine at -5 °C. *Environ Microbiol.* 13: 2269–2278.

Baldini RL, Avedissian M, Gomes SL. 1998. The CIRCE element and its putative repressor control cell cycle expression of the *Caulobacter crescentus groESL* operon. *J Bacteriol.* 180: 1632-41.

Basak S, Banerjee T, Gupta SK, Ghosh TC. 2004. Investigation on the Causes of Codon and Amino Acid usages variation between Thermophilic *Aquifex aeolicus* and Mesophilic *Bacillus subtilis. J Biomol Struct Dyn*. 22: 205-14.

Basak S, Ghosh TC. 2006. Temperature adaptation of synonymous codon usage in different functional categories of genes: a comparative study between homologous genes of *Methanococcus jannaschii* and *Methanococcus maripaludis. FEBS Lett*. 580: 3895-9.

Bennetzen JL, Hall BD. 1982. Codon selection in yeast. *J Biol Chem.* 257: 3026-31.

Berezovsky IN, Shakhnovich EI. 2005. Physics and evolution of thermophilic adaptation. *Proc Natl Acad Sci U S A*. 102: 12742-12747.

Berg JM, Tymoczko JL, Stryer L. 2002. Chemical Bonds in Biochemistry. Biochemistry (5th ed.) W. H. Freeman, New York.

Blöchl E, Rachel R, Burggraf S, Hafenbradl D, Jannasch HW, Stetter KO. 1997. *Pyrolobus fumarii*, gen and sp nov, represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C. *Extremophiles*. 1: 14–21.

Boisvert DC, Wang J, Otwinowski Z, Horwich AL, Sigler PB. 1996. The 24 A° crystal structure of the bacterial chaperonin GroEL complexed with ATP gamma S. *Nat Struct Biol*. 3: 170-7.

Braig K, Adams PD, Brunger AT. 1995. Conformational variability in the refined structure of the chaperonin GroEL at 28 Å resolution. *Nat Struct Biol*. 2: 1083–1094.

Braig K, Otwinowski Z, Hegde R, Boisvert DC, Joachimiak A, Horwich AL, Sigler PB. 1994. The crystal structure of the bacterial chaperonin GroEL at 28 A°. *Nature*. 371: 578–586.

Bronnenmeier K, Kern A, Liebl W, Staudenbauer WL. 1995. Purification of *Thermotoga maritima* enzymes for the degradation of cellulosic materials. *Appl Environ Microbiol*. 61: 1399-407.

Bucca G, Hindle Z, Smith CP. 1997. Regulation of the *dnaK* operon of *Streptomyces coelicolor* A3(2) is governed by HspR, an autoregulatory repressor protein. *J Bacteriol.* 179: 5999-6004.

Bukau B, Horwich AL. 1998. The Hsp70 and Hsp60 chaperone machines. *Cell.* 92: 351–366.

Cary SC, McDonald IR, Barrett JE, Cowan DA. 2010. On the rocks: the microbiology of Antarctic Dry Valley soils. *Nat Rev Microbiol*. 8: 129-38.

Casanueva A, Tuffin M, Cary C, Cowan DA. 2010. Molecular adaptations to psychrophily: the impact of 'omic' technologies. *Trends Microbiol.* 18: 374-81.

Catanzano F, Graziano G, Capasso S, Barone G. 1997. Thermodynamic analysis of the effect of selective monodeamidation at asparagine 67 in ribonuclease A. *Protein Sci*. 6: 1682-1693.

Chakravarty S, Varadarajan R. 2000. Elucidation of determinants of protein stability through genome sequence analysis. *FEBS*. 470: 65-69.

Chang BY, Chen KY, Wen YD, Liao CT. 1994. The response of a *Bacillus subtilis* temperature-sensitive *sigA* mutant to heat stress. *J Bacteriol*. 176: 3102–3110.

Chintalapati S, Kiran MD, Shivaji S. 2004. Role of membrane lipid fatty acids in cold adaptation. *Cell Mol Biol*. 50: 631–642.

Chou KC. 1992. Energy-optimized structure of antifreeze protein and its binding mechanism. *J Mol Biol*. 223: 509–517.

Chung YC, Kobayashi T, Kanai H, Akiba T, Kudo T. 1995. Purification and properties of extracellular amylase from the hyperthermophilic archeon *Thermococccus profundus* DT5432. *Appl Environ Microbiol*. 61: 1502-1506

Ciaramella M, Cannio R, Moracci M, Pisani FM, Rossi M. 1995. Molecular biology of extremophiles. *World J Microbiol Biotechnol*. 11: 71-84.

Clary DO, Wolstenholme DR. 1985. The mitochondrial DNA molecular of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J Mol Evol.* 22: 252-71.

Collins T, Meuwis MA, Gerday C, Feller G. 2003. Activity, stability and flexibility in glycosidases adapted to extreme thermal environments. *J Mol Biol*. 328: 419–428.

Comeron JM. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J Mol Evol.* 41: 1152-9.

Cook GM, Russell JB, Reichert A, Wiegel J. 1996. The intracellular pH of *Clostridium paradoxum,* an anaerobic, alkaliphilic, and thermophilic bacterium. *Appl. Environ. Microbiol.* 62: 4576–4579.

Coyle JE, Jaeger J, Gross M, Robinson CV, Radford SE. 1997. Structural and mechanistic consequences of polypeptide binding by GroEL. *Fold Des*. 2: R93-104.

Craig EA, Gambill BD, Nelson RJ. 1993. Heat shock proteins: molecular chaperones of protein biogenesis. *Microbiol Rev*. 57: 402-414.

Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol Biol Evol.* 16: 372-82.

Creighton TE. 1997. Proteins: structures and molecular properties. Freeman & Company, New York.

Cross SJ, Ciruela A, Poomputsa K, Romaniec MP, Freedman RB. 1996. Thermostable chaperonin from *Clostridium thermocellum*. *Biochem J*. 316: 615-22.

D'Amico S, Collins T, Marx JC, Feller G, Gerday C. 2006. Psychrophilic microorganisms: challenges for life. *EMBO Rep*. 7: 385-9.

D'Amico S, Marx JC, Gerday C, Feller G. 2003. Activity–stability relationships in extremophilic enzymes. *J Biol Chem*. 278: 7891–7896.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*. 5: 345-352.

De Maio A. 1999. Heat shock proteins: facts, thoughts, and dreams. *Shock.* 11: 1-12.

Deming JW. 2002. Psychrophiles and polar regions. *Current Opinion in Microbiology.* 5: 301–309.

Denisov VP, Venu K, Peters J, Hörlein HD, Halle B. 1997. Orientational disorder and entropy of water in protein cavities. *J Phys Chem.* 101: 9380-9389.

Duckworth AW, Grant WD, Jones BE, Vansteenbergen R. 1996. Phylogenetic diversity of soda lake alkaliphiles. *FEMS Microbiol. Ecol.* 19: 181–191

Eastman JT. 1993. Antarctic Fish Biology: Evolution in a Unique Environment, Academic Press, San Diego, Calif, USA, 1993.

Eaton T, Shearman C, Gasson M. 1993. Cloning and sequence analysis of the *dnaK* gene region of *Lactococcus lactis* subsp *lactis. J Gen Microbiol*. 139: 3253–3264.

Eggen HIL, Geerling A, Watts J, Vos WM de. 1990. Characterization of pyrolysin, a hyperthermoactive serine protease from the archaeobacterium *Pyrococcus furiosus*. *FEMS Microbiol Lett.* 71: 17-20.

Ellis RJ (Ed). The chaperonins. san Diego: Academic Press; 1996.

Endo T, Ikeo K, Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Mol Biol Evol.* 13: 685-690.

Ermolaeva MD. 2001. Synonymous codon usage in bacteria. *Curr Issues Mol Biol.* 3: 91-97.

Feller G. 2013. Psychrophillic enzymes: from folding to function and biotechnology. *Scientifica*. 2013: 1- 28.

Fenton WA, Horwich AL. 1997. GroEL-mediated protein folding. *Protein Sci.* 6: 743-760.

Fenton WA, Kashi Y, Furtak K, Norwich AL. 1994. Residues in chaperonin GroEL required for polypeptide binding and release. *Nature*. 371: 614-619.

Ferrer M, Chernikova TN, Yakimov MM, Golyshin PN, Timmis KN. 2003. Chaperonins govern growth of *Escherichia coli* at low temperatures. *Nat Biotech*. 21: 1266-1267.

Forterre, P. Strategies of hyperthermophiles in nucleic acids adaptation to high temperature. In: Unesco Encyclopedia of Life Support Systems, Theme 6.73: Extremophiles, in press.

Fox JM, Erill I. 2010. Relative codon adaptation: a generic codon bias index for prediction of gene expression. *DNA Res*. 17: 185-96.

French S, Robson B. 1983. What is a conservative substitution? *J Mol Evol*. 19: 171-175.

Friedmann EI. 1982. Endolithic microorganisms in the Antarctic cold desert. *Science*. 215: 1045–1053.

Friedrich A, Antranikian G. 1996. Keratin degradation by *Fervidobacterium pennavorans*, a novel thermophilic anaerobic species of the order Thermotogales. *Appl Environ Microbiol.* 62: 2875-2882

Fujita M, Kanehisa M. 2005. Comparative Analysis of DNA-Binding Proteins between Thermophilic and Mesophilic Bacteria. *Genome Informatics*. 16: 174-181.

Fukuchi S, Nishikawa K. 2001. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J Mol Biol.* 309: 835–843.

Galtier N, Lobry JR. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol*. 44: 632-636.

Gamer J, Multhaup G, Tomoyasu T, McCarty JS, Rudiger S, Schonfeld HJ, Schirra C, Bujard H, Bukau B. 1996. A cycle of binding and release of the *DnaK, DnaJ* and *GrpE* chaperones regulate activity of the *Escherichia coli* heat shock transcription factor sigma 32. *EMBO J*. 15: 607–617.

Garcia-Vallvé S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res*. 10: 1719–1725.

Giordano D, Russo R, di Prisco G, Verde C. 2012. Molecular adaptations in Antarctic fish and marine microorganisms. *Marine Genomics.* 6: 1–6.

Goldberg AL. 2003. Protein degradation and protection against misfolded or damaged proteins. *Nature.* 426: 895-9.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11: 725-36.

Goldstein RA. 2007. Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: insights from the quasi-chemical approximation. *Protein Sci*. 16: 1887-95.

Goloubinoff P, Christeller JT, Gatenby AA, Lorimer GH. 1989. Reconstitution of active dimeric ribulose bisphosphate carboxylase from an unfolded state depends on two chaperonin proteins and Mg-ATP. *Nature*. 342: 884-889.

Gomes SL, Gober JW, Shapiro L. 1990. Expression of the *Caulobacter* heat shock gene *dnaK* is developmentally controlled during growth at normal temperatures. *J Bacteriol*. 172: 3051–3059.

Gorgan DW. 1998. Hyperthermophiles and the problem of DNA instability. *Molec Microbiol*. 28: 1043–1049.

Gorovits BM, Ybarra J, Seale JW, Horowitz PM. 1997. Conditions for nucleotide-dependent GroES-GroEL interactions. *J Biol Chem.* 272: 26999-27004.

Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res*. 10: 7055-7074.

Gragerov A, Nudler E, Komissarova N, Gaitanaris GA, Gottesman ME, Nikiforov V. 1992. Cooperation of GroEL/GroES and DnaK/DnaJ heat shock proteins in preventing protein misfolding in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 89: 10341-4.

Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. 1980a. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res*. 8: r49–r62.

Gromiha MM, Oobatake M, Sarai A. 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys Chem*. 82: 51-67.

Gromiha MM, Suresh MX. 2008. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins*. 70: 1274-9.

Groth I, Schumann P, Rainey FA, Martin K, Schuetze B, Augsten K. 1997. *Bogoriella caseilytica* gen. nov., sp. nov., a new alkaliphilic actinomycete from a soda lake in Africa. *Int. J. Syst. Bacteriol*. 47:788–794.

Gunesekere IC, Kahler CM, Powell DR, Snyder LA, Saunders NJ, Rood JI, Davies JK. 2006. Comparison of the RpoH-Dependent Regulon and General Stress Response in *Neisseria gonorrhoeae. J Bacteriol*. 188: 4769–4776.

Gupta SK, Bhattacharyya TK, Ghosh TC. 2004. Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. *J Biomol Struct Dyn*. 21: 527–536.

Hansen JE, Gafni A. 1993. Thermal switching between enhanced and arrested reactivation of bacterial glucose-6-phosphate dehydrogenase assisted by GroEL the absence of ATP. *J Biol Chem*. 268: 21632-21636.

Hao W, Golding GB. 2004. Patterns of bacterial gene movement. *Mol Biol Evol*. 21: 1294–1307.

Harrison CJ, Hayer-Hartl M, Di Liberto M, Hartl F, Kuriyan J. 1997. Crystal structure of the nucleotide exchange factor GrpE bound to the *ATPase* domain of the molecular chaperone DnaK. *Science*. 276: 431–435.

Hartl FU, Hayer-Hartl M. 2002. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*. 295: 1852–1858.

Hayashi K, Nimura Y, Ohara N, Uchimura T, Suzuki H, Komagata K, Kozaki M. 1996. Low-temperature active cellulase produced by *Acremonium alcalophilum* JCM 7366. *Seibutsu-kogaku Kaishi*. 74: 7-10.

Hayer-Hartl MK, Weber F, Hartl FU. 1996. Mechanism of chaperonin action: GroES binding and release can drive GroEL-mediated protein folding in the absence of ATP hydrolysis. *EMBO J*. 15: 6111-6121.

Hecker M, Schumann W, Völker V. 1996. Heat-shock and general stress response in *Bacillus subtilis. Mol Microbiol*. 19: 417–428.

Hendrick JP, Hartl FU. 1993. Molecular chaperone functions of heat-shock proteins. *Annu Rev Biochem*. 62: 349-384.

Herman C, Thévenet D, D'Ari R, Bouloc P. 1995. Degradation of ς^{32}, the heat shock regulator in *Escherichia coli*, is governed by HflB. *Proc Natl Acad Sci USA*. 92: 3516–3520.

Hershberg R, Petrov DA. 2008. Selection on Codon Bias. *Annu Rev Genet*. 42: 287–99.

Hjerde E, Lorentzen MS, Holden MT, Seeger K, Paulsen S, Bason N, Churcher C, Harris D, Norbertczak H, Quail MA, Sanders S, Thurston S, Parkhill J, Willassen NP, Thomson NR. 2008. The genome sequence of the fish pathogen *Aliivibrio salmonicida* strain LFI1238 shows extensive evidence of gene decay. *BMC Genomics*. 9: 616.

Hogan C. 2012. Extremophile. Retrieved from http://www.eoearth.org/view/article/51cbf01c7896bb431f6a0508

Homuth G, Masuda S, Mogk A, Kobayashi Y, Schumann W. 1997. The dnaK operon of *Bacillus subtilis* is heptacistronic. *J Bacteriol.* 179: 1153-64.

Horikoshi K. 1991. Microorganisms in alkaline environments. (Kodansha-VCH, Tokyo, Japan).

Horwich AL, Brooks Low K, Fenton WA, Hirshfield IN, Furtak K. 1993. Folding in vivo of bacterial cytoplasmic proteins: role of GroEL. *Cell*. 74: 909–917.

Hou ZC, Yang N. 2003. Factors Affecting Codon Usage in *Yersinia pestis. ACTA Biochimica Et Biophysica Sinica*. 35: 580-586.

Houry WA, Frishman D, Eckerskorn C, Lottspeich F, Hartl FU. 1999. Identification of *in vivo* substrates of the chaperonin GroEL. *Nature*. 402: 147–154.

Huber R, Huber H, Stetter KO. 2000. Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. *FEMS Microbiol Rev*. 24: 615-623.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals over dominant selection. *Nature*. 335: 167-70.

Hunt JF, Weaver AJ, Landry SJ, Gierasch L, Deisenhofer J. 1996. The crystal structure of the GroES co-chaperonin at 28A° resolution. *Nature*. 379: 37–45.

Ikeda M, Clark SC. 1998. Molecular cloning of extremely thermostable esterase gene from hyperthermophilic archeon *Pyrococcus furiosus* in *Escherichia coli. Biotechnol Bioeng*. 57: 624-629.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*. 2: 13-34.

Ikemura. T. 1981a. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol.* 146: 1-21.

Ikemura. T. 1981b. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol*. 151: 389-409.

Ikemura. T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol*. 158: 573-97.

Ina Y.1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J Mol Evol.* 40: 190-226.

Ingvarsson PK. 2008. Molecular evolution of synonymous codon usage in *Populus*. *BMC Evol Biol*. 8: 307-319.

Inobe T, Makio T, Takasu-Ishikawa E, Terada TP, Kuwajima K. 2001. Nucleotide binding to the chaperonin GroEL: non-cooperative binding of ATP analogs and ADP, and cooperative effect of ATP. *Biochim Biophys Acta.* 1545: 160-173.

Jaenicke R. 1991. Protein stability and molecular adaptation to extreme conditions. *Eur J Biochem*. 202: 715-728.

Jaenicke R. 2000. Stability and stabilization of globular proteins in solution. *J Biotechnol*. 79: 193-203.

Jäger S, Jäger A, Klug G. 2004. CIRCE is not involved in heat-dependent transcription of *groESL* but in stabilization of the mRNA 5′-end in *Rhodobacter capsulatus. Nucleic Acids Res*. 32: 386-396.

Jahandideh M, Barkooie SM, Jahandideh S, Abdolmaleki P, Movahedi MM, Hoseini S, Asadabadi EB, Jouni FJ, Karami Z, Firoozabadi NH. 2008. Elucidating the protein cold-adaptation: Investigation of the parameters enhancing protein psychrophilicity. *J Theor Biol*. 255: 113-8.

Jia Z, Davies PL. 2002. Antifreeze proteins: an unusual receptor–ligand interaction. *Trends Biochem Sci*. 27: 101–106.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. Pp. 2 1-132 *in* h. N. Munro, ed. Mammalian protein metabolism iii. Academic press, New York.

Kafatos FC, Efstratiadis A, Forget BG, Weissman SM. 1977. Molecular evolution of human and rabbit beta-globin mRNAs. *Proc Natl Acad Sci U S A*. 74: 5618–5622.

Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene.* 238: 143-155.

Kannan N, Vishveshwara S. 2000. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. *Protein Eng*. 13: 753-761.

Kashefi K, Lovley DR. 2003. Extending the upper temperature limit for life. *Science*. 15: 934.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW. 2007. Patterns and Implications of Gene Gain and Loss in the Evolution of *Prochlorococcus. PLoS Genet*. 3: e231.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*. 267: 275-276.

Koch R, Spreinat K, Lemke K, Antranikian G. 1991. Purification and properties of a hyperthermoactive a-amylase from the archaeobaterium *Pyrococcus woesei. Arch Microbiol.* 155: 572 - 578.

Korber B. 2000. HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences, Chapter 4, pages 55-72. Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Publishers.

Kreil DP, Ouzounis CA. 2001. Identification of thermophilic species by amino acid compositions deduced from their genomes. *Nucliec Acid Res.* 29: 1608-1615.

Kube M, Migdoll AM, Gehring I, Heitmann K, Mayer Y, Kuhl H, Knaust F, Geider K, Reinhardt R. 2010. Genome comparison of the epiphytic bacteria *Erwinia billingiae* and *E. tasmaniensis* with the pear pathogen *E. pyrifoliae. BMC Genomics*. 11: 393.

Kumar R, Prabhu NP, Rao DK, Bhuyan AK. 2006. The Alkali Molten Globule State of Horse Ferricytochrome c: Observation of Cold Denaturation. *J Mol Biol*. 364: 483-495.

Kumar S, Nussinov R. 2001. How do thermophilic proteins deal with heat? *Cell Mol Life Sci*. 58: 1216-1233.

Kumar S, Tsai C-J, Nussinov R. 2000. Factors enhancing protein thermostability. *Protein Eng*. 13: 179-191.

Kunte HJ, Truper HG, Stan-Lotter H. 2002. Halophilic microorganisms, p. 185-200. In Horneck G, Baumstark-Khan C (ed.), Astrobiology, the quest for the conditions of life. Springer, Koln, Germany.

Kwak YS, Akeba T, Kudo T. 1998. Purification and characterization from hyperthermophilic archaeon *Thermococcus profundus*, which hydrolyzes both a-1,4 and a-1,6 glucosidic linkages. *J Ferment Bioeng*. 86: 363-367.

Lafay B, Atherton JC, Sharp PM. 2000. Absence of translationally selected synonymous codon usage bias in Helicobacter pylori. *Microbiology*. 146: 851-60.

Lao PJ, Forsdyke DR. 2000. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res*. 10: 228-236.

Laskos L, Ryan CS, Fyfe JA, Davies JK. 2004. The RpoH mediated stress response in *Neisseria gonorrhoeae* is regulated at the level of activity. *J Bacteriol*. 186: 8443–8452.

Lebbink JHG. 1999. Molecular characterisation of the thermostability and catalytic properties of enzymes from hyperthermophiles PhD Thesis Wageningen University, Wageningen.

Lee JT, Kanai H, Kobayashi T, Akiba T, Kudo T. 1996. Cloning, nucleotide sequence and hyperexpression of a-amylase gene from an archaeon, *Thermococcus profundus*. *J Ferment Bioeng*. 82: 432-438.

Lee MC, Marx CJ. 2012. Repeated, Selection-Driven Genome Reduction of Accessory Genes in Experimental Populations. *PLoS Genet*. 8: 1-8.

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and non synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*. 2: 150-74.

Li Z, Srivastava P. 2004. Heat-shock proteins. *Current Protocols in Immunology*. 58: A1T1–A1T6.

Lin Z, Schwarz FP, Eisenstein E. 1995. The Hydrophobic Nature of GroEL-Substrate Binding. *J Biol Chem*. 270: 1011-1014.

Liu J, Xu X, Stormo GD. 2008. The *Cis*-regulatory map of *Shewanella* genomes. *Nucleic Acids Res*. 36: 5376-90.

Lonhienne T, Zoidakis J, Vorgias CE, Feller G, Gerday C, Bouriotis V. 2001. Modular structure, local flexibility and cold-activity of a novel chitobiase from a psychrophilic Antarctic bacterium. *J Mol Biol.* 310: 291-7.

Lopez-Camacho C, Salgado J, Lequerica JL, Madarro A, Ballestar E, Franco L, Polaina J. 1996. Amino acid substitutions enhancing thermostability of *Bacillus polymyxa* beta-glucosidase A. *Biochem J*. 314: 833-838.

Low PS, Bada JL, Somero GN. 1973. Temperature adaptation of enzymes: roles of the free energy, the enthalpy, and the entropy of activation. *Proc Natl Acad Sci U S A*. 70: 430–432.

Lynn DJ, Singer GA, Hickey DA. 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res*. 30: 4272-7.

Ma JC, Dougherty DA. 1997. The cation-pi interaction. *Chem Rev*. 97: 1303-1324.

Madigan MT, Martinko JM, Parker J. 1997. In Brock Biology of Microorganisms Madigan MT, Martinko JM & Parker J (eds) Upper Saddle River, NJ, USA: Prentice Hall International Editions.

Madigan MT, Martino JM. 2006. *Brock Biology of Microorganisms* (11th ed.). Pearson. p. 136.

Manakova E. 2001. Study of structure and function of GroELS chaperonin system using small angle scattering. Dissertation, TU München.

Matthews BW, Nicholson H, Becktel WJ. 1987. Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc Natl Acad Sci USA*. 84: 6663–6667.

Mattos C. 2002. Protein-interactions in a dynamic world. *Trends Biochem Sci*. 27: 203-208.

Mayhew M, da Silva ACR, Martin J, Erdjument-Bromage H, Tempst P, Hartl FU. 1996. Protein folding in the central cavity of the GroEL-GroES chaperonin complex. *Nature*. 379: 420–426.

McDonald JH, Grasso AM, Rejto LK. 1999. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. *Mol Biol Evol.* 16: 1785-90.

McGuire AM, Hughes JD, Church GM, 2000. Conservation of DNA Regulatory Motifs and Discovery of New Motifs in Microbial Genomes. *Genome Res*. 10: 744-757.

Médigue C, Krin E, Pascal G, Barbe V, Bernsel A, Bertin PN, Cheung F, Cruveiller S, D'Amico S, Duilio A, Fang G, Feller G, Ho C, Mangenot S, Marino G, Nilsson J, Parrilli E, Rocha EP, Rouy Z, Sekowska A, Tutino ML, Vallenet D, von Heijne G, Danchin A. 2005. Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. *Genome Res.*15: 1325-35.

Metpally RP, Reddy BV. 2009. Comparative proteome analysis of psychrophilic versus mesophilic bacterial species: Insights into the molecular basis of cold adaptation of proteins. *BMC Genomics*. 10: 11.

Miller W, Makova KD, Nekrutenko A, Hardison RC. 2004. Comparative genomics. *Annu Rev Genomics Hum Genet*. 5: 15–56.

Mirkin BG, Fenner TI, Galperin MY, Koonin EV. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*. 3: 2.

Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol*. 16: 23-36.

Mueller SW. 1947. Permafrost or permanently frozen ground and related engineering problems United States Army, Washington, DC.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. 11: 715-24.

Nagendra HG, Sukumar N, Vijayan M. 1998. Role of water in plasticity, stability, and action of proteins: the crystal structures of lysozyme at very low levels of hydration. *Proteins*. 32: 229-240.

Nakamura T, Tanaka M, Maruyama A, Higashi Y, Kurusu Y. 2004. A Nonconserved Carboxy-Terminal Segment of GroEL Contributes to Reaction Temperature. *Biosci Biotechnol Biochem*. 68: 2498-2504.

Narberhaus F, Bahl H. 1992. Cloning, sequencing, and molecular analysis of the *groESL* operon of *Clostridium acetobutylicum. J Bacteriol*. 174: 3282–3289.

Nei M, Gojobori T. 1986. Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions. *Mol Biol Evol*. 3: 418-426.

Niehaus F, Bertoldo C, Kähler M, Antranikian G. 1999. Extremophiles as a source of novel enzymes for industrial application. *Appl Microbiol Biotechnol*. 51: 711-29.

Oehler S, Eismann E R, Krämer H, Müller-Hill B. 1990. The three operators of the *lac* operon cooperate in repression. *EMBO J*. 9: 973–979.

Ohta T, Saito K, Kuroda M, Honda K, Hirata H, Hayashi H. 1994. Molecular cloning of two new heat shock genes related to the *hsp70* genes in *Staphylococcus aureus. J Bacteriol*. 176: 4779–4783.

Osváth A, Szilágyi J, Kardos S, Barna L, Závodszky P. 2007. "10 Protein Folding." *Handbook of Neurochemistry and Molecular Neurobiology: Neural Protein Metabolism and Function* 3.

Ota T, Nei M. 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. *Mol Biol Evol*. 11: 613-9.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*. 96: 2896-901.

Pack SP, Yoo YJ. 2004. Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. *J Biotechnol*. 111: 269-77.

Pamilo P, Bianchi NO. 1993. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol Biol Evol*. 10: 271-81.

Panasik N, Brenchley JE, Farber GK. 2000. Distributions of structural features contributing to thermostability in mesophilic and thermophilic alpha/beta barrel glycosyl hydrolases. *Biochim Biophys Acta*. 1543: 189-201.

Pandey KD, Shukla SP, Shukla PN, Giri DD, Singh JS, Singh P, Kashyap AK. 2004. Cyanobacteria in Antarctica: ecology, physiology and cold adaptation. *Cell Mol Biol*. 50: 575–584.

Parsell DA, Lindquist S. 1993. The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins. *Ann Rev Genet*. 27: 437-496.

Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner R, Dodgson J. 1980. The evolution of genes: the chicken preproinsulin gene. *Cell*. 20: 555-66.

Phadtare S. 2004. Recent developments in bacterial cold-shock response. *Curr Issues Mol Biol*. 6: 125–136.

Post LE, Nomura M. 1980. DNA sequences from the *str* operon of *Escherichia coli*. *J Biol Chem*. 255: 4660-6.

Pritsker M, Liu YC, Beer MA, Tavazoie S. 2004. Whole-Genome Discovery of Transcription Factor Binding Sites by Network-Level Conservation. *Genome Res*. 14: 99–108.

Querol E, Perez-Pons JA, Mozo-Villarias A. 1996. Analysis of protein conformational characteristics related to thermostability. *Protein Engineering*. 9: 265-271.

Rampelotto, PH. 2013. Extremophiles and Extreme Environments. *Life.* 3: 482-485.

Ranjan A, Vidyarthi AS, Poddar R. 2007. Evaluation of codon bias perspectives in phage therapy of *Mycobacterium tuberculosis* by multivariate analysis. *In Silico Biology*. 7: 423–431.

Roberts RC, Toochinda C, Avedissian M, Baldini RL, Gomes SL, Shapiro L. 1996. Identification of a *Caulobacter crescentus* operon encoding *hrcA*, involved in negatively regulating heat-inducible transcription, and the chaperone gene *grpE. J Bacteriol*. 178: 1829–1841.

Rodrigues DF, Tiedje JM. 2008. Coping with Our Cold Planet. *Appl Environ Microbiol.* 74: 1677-86.

Rohde R, Price PB. 2007. Diffusion-controlled metabolism for long-term survival of single isolated microorganisms trapped within ice crystals. *Proc Natl Acad Sci U S A*. 104: 16592-16597.

Roman Dial C, Dial RJ, Saunders R, Lang SA, Lee B, Wimberger P, Dinapoli MS, Egiazarov AS, Gipple SL, Maghirang MR, Swartley-McArdle DJ, Yudkovitz SR, Shain DH. 2012. Historical biogeography of the North American glacier ice worm, *Mesenchytraeus solifugus* (Annelida: Oligochaeta: Enchytraeidae). *Mol Phylogenet Evol*. 63: 577–584.

Romero PA, Arnold FH. 2009. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol.* 10: 866-76.

Rosen R, Ron EZ. 2002. Proteome analysis in the study of the bacterial heat-shock response. *Mass Spectrom Rev*. 21: 244–265.

Russell NJ. 1997. Psychrophilic bacteria molecular adaptations of membrane lipids. *Comp Biochem Physiol Physiol*. 118: 489–493.

Russell RJ, Hough DW, Danson MJ, Taylor GL. 1994. The crystal structure of citrate synthase from the thermophilic archaeon *Thermoplasma acidophilum*. *Structure*. 2: 1157-1167.

Rye HS, Burston SG, Fenton WA, Beechem JM, Xu Z, Sigler PB, Horwich AL. 1997. Distinct actions of *cis* and *trans* ATP within the double ring of the chaperonin GroEL. *Nature*. 388: 792-798.

Rye HS, Roseman AM, Chen S, Furtak K, Fenton WA, Saibil HR, Horwich AL. 1999. GroEL-GroES cycling: ATP and nonnative polypeptide direct alternation of folding-active rings. *Cell*. 97: 325–338.

Sadeghi M, Naderi-Manesh H, Zarrabi M, Ranjbar B. 2006. Effective factors in thermostability of thermophilic proteins. *Biophys Chem*. 119: 256-270.

Sakikawa C, Taguchi H, Makino Y, Yoshida M. 1999. On the maximum size of proteins to stay and fold in the cavity of GroEL underneath GroES. *J Biol Chem*. 274: 21251–21256.

Sattler B, Puxbaum H, Psenner R. 2001. Bacterial growth in supercooled cloud droplets. *Geophysical Research Letters.* 28: 239–242.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR, Post AF, Hagemann M, Paulsen I, Partensky F. 2009. Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev*. 73: 249-99.

Schröder H, Langer T, Hartl F-U, Bukau B. 1993. *DnaK, DnaJ, GrpE* form a cellular chaperone machinery capable of repairing heat-induced protein damage. *EMBO J*. 12: 4137–4144.

Schulz A, Schumann W, 1996. *HrcA*, the first gene of the *Bacillus subtilis dnaK* operon encodes a negative regulator of class I heat shock genes. *J Bacteriol*. 178: 1088-1093.

Schulz A, Tzschaschel B, Schumann W. 1995. Isolation and analysis of mutants of the *dnaK* operon of *Bacillus subtilis*. *Mol Microbiol*. 15: 421-9.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res*. 33: 1141-1153.

Sharp PM, Li WH. 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol*. 24: 28-38.

Sharp PM, Li WH. 1987. The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 15: 1281-1295.

Sharp PM. 1997. In search of molecular Darwinism. *Nature*. 385: 111–112.

Shewmaker F, Maskos K, Simmerling C, Landry SJ. 2001. The Disordered Mobile Loop of GroES Folds into a Defined β-Hairpin upon Binding GroEL. *J Biol Chem*. 276: 31257-31264.

Shimamura T, Koike-Takeshita A, Yokoyama K, Masui R, Murai N, Yoshida M, Taguchi H, Iwata S. 2004. Crystal Structure of the Native Chaperonin Complex from *Thermus thermophilus* Revealed Unexpected Asymmetry at the *cis*-Cavity. *Structure*. 12: 1471-1480.

Sigler PB, Xu Z, Rye HS, Burston SG, Fenton WA, Horwich AL. 1998. Structure and function in GroEL-mediated protein folding. *Annu Rev Biochem*. 67: 581–608.

Singer GA, Hickey DA. 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*. 317: 39-47.

Smith SA, Beaulieu JM, Donoghue MJ. 2009. MEGA 50-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol Biol*. 9: 37-48.

Spivakov M, Akhtar J, Kheradpour P, Beal K, Girardot C, Koscielny G, Herrero J, Kellis M, Furlong EE, Birney E. 2012. Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biol*. 13: R49.

Sterner R, Liebl W. 2001. Thermophilic Adaptation of Proteins. *Crit. Rev. Biochem. Mol. Biol*. 36: 39-106.

Stetter, KO. 1986. in: T.D. Brock (Ed.) *Thermophiles: General, molecular and applied microbiology*, John Wiley & Sons, Inc., New York. 39.

Stetter KO. 1996. Hyperthermophiles in the history of life. *Ciba Found Symp*. 202: 1-10.

Szabo A, Langer T, Schroder H, Flanagan J, Bukau B, Hartl FU. 1994. The ATP hydrolysis-dependent reaction cycle of the *Escherichia* *coli* Hsp70 system *DnaK*, *DnaJ* and *GrpE. Proc Natl Acad Sci* USA. 91: 10345–10349.

Szilágyi A, Závodszky P. 2000. Structural differences between mesophilic moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure*. 8: 493-504.

Takai K, Nakamura K, Toki T, Tsunogai U, Miyazaki M, Miyazaki J, Hirayama H, Nakagawa S, Nunoura T, Horikoshi K. 2008. Cell proliferation at 122 degrees C and isotopically heavy CH4 production by a hyperthermophilic methanogen under high-pressure cultivation. *Proc Natl Acad Sci*. 105: 10949-54.

Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*. 9: 678-687.

Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo GD. 2001. A comparative genomics approach to prediction of new members of regulons. *Genome Res*. 11: 566-584.

Tanabe H, Kobayashi Y, Akamatsu I. 1988. Pretreatment of pectic wastewater with pectate lyase from an alkalophilic *Bacillus sp*. *Agric. Biol. Chem.* 52: 1855–1856.

Tanaka Y, Tsumoto K, Yasutake Y, Umetsu M, Yao M, Fukada H, Tanaka I, Kumagai I. 2004. How oligomerization contributes to the thermostability of an archaeon protein Protein l-isoaspartyl-O-methyltransferase from *Sulfolobus tokodaii. J Biol Chem.* 279: 32957-32967.

Tanner JJ, Hecht RM, Krause KL. 1996. Determinants of Enzyme Thermostability Observed in the Molecular Structure of *Thermus aquaticus* d-Glyceraldehyde-3-phosphate Dehydrogenase at 25 Å Resolution. *Biochemistry.* 35: 2597-2609.

Thirumalai D, Lorimer GH. 2001. Chaperonin-mediated protein folding. *Annu Rev Biophys Biomol Struct.* 30: 245-269.

Thompson JD, Higgins DG Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.

Tomoyasu T, Gamer J, Bukau B, Kanemori M, Mori H, Rutman AJ, Oppenheim AB, Yura T, Yamanaka K, Niki H, et al. 1995. *Escherichia coli* FtsH is a membrane-bound, ATP-dependent protease which degrades the heat-shock transcription factor sigma 32. *EMBO J.* 14: 2551-60.

Trivedi S, Satyawada RR, Hukam SG. 2005. Nucleic acid stability in thermophilic prokaryotes: a review. *Journal of Cell and Molecular Biology*. 4: 61-69.

Valpuesta JM, Martín-Benito J, Gómez-Puertas P, Carrascosa JL, Willison KR. 2002. Structure and function of a protein folding machine: the eukaryotic chaperonin CCT. *FEBS Lett.* 529: 11-16.

Veltman OR, Vriend G, Middelhoven PJ, Van Den Burg B, Venema G, Eijsink VGH. 1996. Analysis of structural determinants of the stability of thermolysin like proteases by molecular modelling and site-directed mutagenesis. *Protein Eng* 9: 181-1189.

Vezzi A, Campanaro S, D'Angelo M, Simonato F, Vitulo N, Lauro FM, Cestaro A, Malacrida G, Simionati B, Cannata N, Romualdi C, Bartlett DH, Valle G. 2005. Life at depth: *Photobacterium profundum* genome sequence and expression analysis. *Science.* 307: 1459-61.

Vieille C, Zeikus GJ. 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev*. 65: 1-43.

Viitanen PV, Gatenby AA, Lorimer GH. 1992. Purified GroEL interacts with the non-native states of a multitude of *E. coli* proteins. *Protein Sci.* 1: 361-369.

Violot S, Aghajari N, Czjzek M, Feller G, Sonan GK, Gouet P, Gerday C, Haser R, Receveur-Brechot V. 2005. Structure of a full length psychrophilic cellulase from *Pseudoalteromonas haloplanktis* revealed by X-ray diffraction and small angle X-ray scattering. *J Mol Biol.* 348: 1211-1224.

Vishniac H, Klinger J. 1986. Yeasts in the Antarctic deserts. In Megusar F, Gantar M (eds) Perspectives in Microbial Ecology. Proceedings of the 4th ISME p 46–51. Ljubljana, Slovenia: Slovene Society for Microbiology.

Vishnivetskaya TA, Petrova MA, Urbance J, Ponder M, Moyer CL, Gilichinsky DA, Tiedje JM. 2006. Bacterial community in ancient Siberian permafrost as characterized by culture and culture-independent methods. *Astrobiology*. 6: 400-414.

Vogt G, Argos P. 1997. Protein thermal stability: hydrogen bonds or internal packing? *Fold Des*. 2: S40-6.

Watanabe K, Hata Y, Kizaki H, Katsube Y, Suzuki Y. 1997. The refined crystal structure of *Bacillus cereus* oligo-1,6-glucosidase at 20 Å resolution: structural characterization of proline-substitution sites for protein thermostabilization. *J Mol Biol*. 269: 142-153.

Weissman JS, Hoh, CM, Kovalenko O, Kashi Y, Chen S, Braig K, Saibil HR, Fenton WA, Horwich AL. 1995. Mechanism of GroEL action: productive release of polypeptide from a sequestered position under GroES. *Cell*. 83: 577–587.

Wels M, Francke C, Kerkhoven R, Kleerebezem M Siezen RJ. 2006. Predicting *cis*-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res*. 34: 1947-1958.

West SE, Iglewski BH. 1988. Codon usage in *Pseudomonas aeruginosa*. *Nucleic Acids Res.* 16: 9323-35.

Wetzstein M, Völker U, Dedio J, Löbau S, Zuber U, Schiesswohl M, Herget C, Hecker M, Schumann W. 1992. Cloning, sequencing, and molecular analysis of the *dnaK* locus from *Bacillus subtilis*. *J Bacteriol*. 174: 3300–3310.

Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, Trinklein ND, Myers RM, Weng Z. 2012. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol* 13: R50.

Willey, Joanne M, Sherwood, Linda, Woolverton, Christopher, J.Prescott, Lansing M. 2008. *Prescott, Harley, and Klein's microbiology/*New York: McGraw-Hill Higher Education.

Wright F. 1990. The 'effective number of codons' used in a gene. *Gene*. 87: 23-9.

Xu Y, Nogi Y, Kato C, Liang Z, Rüger HJ, De Kegel D, Glansdorff N. 2003. *Moritella profunda* sp nov and *Moritella abyssi* sp nov, two psychropiezophilic organisms isolated from deep Atlantic sediments. *Int J Syst Evol Microbiol*. 53: 533-8.

Xu Z, Horwich AL, Sigler PB. 1997. The crystal structure of the asymmetric GroEL–GroES–(ADP)$_7$ chaperonin complex. *Nature*. 388: 741-50.

Yamauchi S, Ueda Y, Matsumoto M, Inoue U, Hayashi H. 2012. Distinct features of protein folding by the GroEL system from a psychrophilic bacterium, *Colwellia psychrerythraea 34H*. *Extremophiles*. 16: 871-82.

Yip KS, Britton KL, Stillman TJ, Lebbink J, de Vos WM, Robb FT, Vetriani C, Maeder D, Rice DW. 1998. Insights into the molecular basis of thermal stability from the analysis of ion-pair networks in the glutamate dehydrogenase family. *Eur J Biochem*. 255: 336-346.

Yu T, Li J, Yang Y, Qi L, Chen B, Zhao F, Bao Q, Wu J. 2012. Codon usage patterns and adaptive evolution of marine unicellular cyanobacteria *Synechococcus* and *Prochlorococcus. Mol Phylogenet Evol.* 62: 206-213.

Yura T, Nagai H, Mori H. 1993. Regulation of the heat-shock response in bacteria. *Annu Rev Microbiol*. 47: 321-350.

Yura T, Yamanaka K, Niki H, Hiraga S, Ogura T. 1995. *Escherichia coli* FtsH is a membrane-bound ATP-dependent protease which degrades the heat-shock transcription factor ς^{32}. *EMBO J.* 14: 2551-2560.

Yuzawa, H, Nagai H, Mori H, Yura T. 1993. Heat induction of sigma 32 synthesis mediated by mRNA secondary structure: a primary step of the heat shock response in *Escherichia coli. Nucleic Acids Res*. 21: 5449–5455.

Zhang G, Fang B. 2006. Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition. *Protein Pept Lett*. 13: 965-970.

Zhou XX, Wang YB, Pan YJ, Li WF. 2008. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids*. 34: 25-33.

Zimmerman SB, Trach SO. 1991. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of *Escherichia coli*. *J Mol Biol*. 222: 599-620.

Zuber U, Schumann W. 1994. CIRCE, a novel heat shock element involved in regulation of heat shock operon *dnaK* of *Bacillus subtilis*. *J Bacteriol*. 176: 1359–1363.

List of publications

1. CyanoPhyChe:A Database for Physico-Chemical Properties, Structure and Biochemical Pathway Information of Cyanobacterial Proteins. V. Parvati Sai Arun[1], Ranjith Kumar Bakku[2,1], **Mranu Subhashini**[1], Pankaj Singh[2], N. Prakash Prabhu[2], Iwane Suzuki[3], Jogadhenu S. S. Prakash[2*]. PLos One. 2012;7(11):e49425.

2. Prediction of cis regulatory elements in the genome of Synechococcus elongatus PCC 6301. P. Parvati Sai Arun, **M. Subhashini,** CH. Santhosh, P. Sankara Krishna, Jogadhenu S. S. Prakash. 2011. Proceedings of the 15th International Congress on Photosynthesis. Photosynthesis Research For Food, Fuel And The Furture. 2013, pp 369-373.

3. Male fertility restoration in transgenic tobacco plants expressing a cysteine protease with plants having targeted expression of cystatin. Pawan shukla, **Mranu Subhashini**, Naveen Singh, Dilip Kumar, Sambasivam Vijayan, Israr Ahmed, Jogadhenu, Prakash, Kirti, P.B (being communicated).

4. Targeted expression of cystatin restores fertility on cysteine protease induced male sterile plants of tobacco. Pawan Shukla; **Mranu**

Subhashini; Naveen K. Singh; Dilip Kumar; S. Vijayan; Israr Ahmed; Jogadhenu S.S. Prakash; P.B. Kirti (being communicated).

List of poster presentations

1. DNA chip as a tool to elucidate functions of hypothetical proteins. **Cell biology conference**- 2009. December 2009. University of Hyderabad.
2. Molecular evolution of GroESL in prokaryotes in terms of adaptation to extreme temperatures. **Indo-German Symposium on Systems Biology** 2012. 27-29 Nov 2012. University of Hyderabad.