Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article **DOI: 10.26479/2018.0406.06**

COMPARATIVE ANALYSIS OF THERMOPHILIC PROTEASES

Atman Vaidya[1], Varun S. Nair[2], John J. Georrge[2*], Singh S. P.[1*]

1.UGC-CAS Department of Biosciences, Saurashtra University, Rajkot, Gujarat, India

2.Department of Bioinformatics, Christ College, Rajkot, Gujarat, India

ABSTRACT: Stability of thermophilic proteases is generally characterized on the basis of their amino acid composition, protein structure, oligomerization, strength interaction, salt bridges and bonding patterns. A single change in Amino Acid sequence may alter the tertiary fold maintaining the protein rigidity. Thermophilic proteases have emerged as highly significant and indispensable part of certain industries such as laundry detergents, pharmaceuticals and food products. Therefore, it becomes necessary to predict the potential stability of thermostable proteases through computational methods to find out the proteins with potential industrial applications. In this study, sequences of thermophilic proteases were retrieved from NCBI Genome and Protein as well as UniProt databases. The seventy-four organisms from which the primary sequence retrieved were classified on the basis of domain, family, genus and temperature. Comparative analysis between thermophilic and mesophilic proteases were performed on the basis of their sequence length, amino acid composition, which are primarily responsible for protein structure, stability and function. The results suggest that alanine and leucine are relatively abundant in thermophiles which contradict the prevailing concept in literature. Further, the protease from *Pyrococcus kukulkanii*, *Thermovibrio ammonificans*, *Thermococcus sibiricus*, *Thermodesulfatator indicus*, *Picrophilus torridus*, *Candidatus Desulfofervidus*, *Candidatus Methanomethylophilus* and *Bellilinea caldifistulae* provided insight into the protein stability through domain and structural analysis. The domain and structural analysis show that most of the proteases are membrane bound and has peptidase M48, Peptidase M50, PDZ and CBS domain. The work can be extended towards *in silico* protein engineering to improve the stability of these thermophilic proteases.

KEYWORDS: Thermophiles, Thermophilic Proteases, Adaptation Mechanisms, Protein Stability, Comparative Studies, Mesophiles and thermophiles

Corresponding Authors: Prof. Dr. Singh S. P.* Ph.D.

UGC-CAS Department of Biosciences, Saurashtra University, Rajkot, Gujarat, India.

Email Address: satyapsingh@yahoo.com, satyapsingh125@gmail.com

Dr. John J. Georrge* Ph.D.

Department of Bioinformatics, Christ College, Rajkot, Gujarat, India.

Email Address: johnjgeorrge@gmail.com

1. INTRODUCTION

Thermophiles, a group of the extremophile, thrive at relatively high temperatures, between 45 and 122 °C (106 and 252 °F) [1]**.** Simple classification of thermophile is distinguished as: Moderate Thermophiles (50-64°C), Extreme Thermophiles (65-79°C) and Hyperthermophiles >80°C [2]**.** An optimal temperature for the existence of hyperthermophiles is above 80°C (176°F). Hyperthermophiles usually belong to the domain Archaea, although some bacteria are also able to tolerate temperatures of around 100°C (212 °F). Some bacteria can live at temperatures higher than 100°C at the depths in sea where water does not boil because of high pressure. Many hyperthermophiles are also able to withstand other environmental extremes such as high acidity or high radiation levels. Hyperthermophiles, a subset of extremophiles [2], are adapted to hot environments for their physiological and nutritional requirements [3]. Despite morphological similarity with bacteria, archaea possess genes and several metabolic pathways that are more closely related to eukaryotes, notably the enzymes involved in transcription and translation. Proteases catalyze the splitting of proteins into smaller peptide fractions and amino acids by the process of proteolysis. Proteases are classified into seven broad groups (Serine proteases, Cysteine proteases, Threonine proteases, Aspartic proteases, Glutamic protease, Metalloproteases and Asparagine peptide lyases) based on their catalytic residues [4]. Many proteases have been isolated and characterized from thermophilic organisms. Thermolysin – a zinc metalloprotease isolated from *Bacillus thermoproteolyticus* is among the most detailed studied protease [5]. Other well characterized thermophilic extracellular proteases include *B. stearothermophilus* neutral proteases 5'6, thermomycolin (from *Malbrancheapulchella* var. *sulfurea),* thermitase (from *Thermoactinomyces vulgaris)* and a protease from *Streptomyces rectus* [6]. Proteases are excreted at relatively low levels by most thermophilic bacteria. However, mesophilic organisms which excrete proteases can do so at activity levels at least one level of magnitude higher. Strict comparisons between thermophilic and mesophilic protease activities are complicated by different assay temperatures and procedures. Production costs and yields would be a less factor if thermophilic proteases were to be considered because of its growth and survival at low ambient temperature [7].The statistical analyses comparing amino acid compositions in mesophilic and thermophilic proteins indicated that the properties most correlated with the proteins of the thermophile include higher residue volume, higher residue hydrophobicity, more charged amino acids (especially Glu, Arg, and Lys), and fewer uncharged polar residues (Ser, Thr, Asn, and Gln) [8].The hydrophobic amino acids content is marginally higher in thermophiles than that in mesophiles and these residues can increase rigidity and hydrophobicity of proteins [9]. A higher Ala content in thermophilic proteins reflects that Ala is the best helix-forming residue (Argos et al., 1979). Although the helices from thermophilic proteins contain a smaller fraction of beta-

branched residues (Val, Ile, and Thr) than helices in mesophilic proteins, and beta-branched residues were found to destabilize a-helix, most systematically analyses showed that thermophilic proteins had higher frequency of Ile and Val compared with mesophilic ones [9]. Among hydrophobic residues, Ala, Val, Leu, and Ile belong to the aliphatic amino acids. It's widely accepted that the aliphatic amino acids would contribute to the hydrophobic interaction, which is main force for maintaining conformational stability in inner part of the protein [10]. Gly is known to contribute to the void volume or cavity in the inner part of the protein structure. Thermophilic proteins have fewer Gly in a particular region of the structure [10]. Pro residue, with their pyrrolidine ring, can only adopt a few configurations and has the lowest conformational entropy, and thus restrict the configurations allowed for the preceding residue. It is known as the residue for making rigid conformation or turn conformation in protein structure [11]. The side chain of Met includes a sulfur atom but remains hydrophobic in nature. Met is known as thermolabile amino acid due to its tendency to undergo oxidation at high temperature. Some systematically analyse reported that thermophilic proteins have lower frequency of Met compared with mesophilic proteins [12]. In order to function at elevated temperatures, the thermophilic proteins must preserve their tertiary folds to maintain their biological function. Certain thermophiles can withstand temperatures above this and have corresponding adaptations to preserve protein function at these temperatures. These can include altered bulk properties of the cell to stabilize all proteins, and specific changes to individual proteins [13]. The presence of salt alters thermostability in the proteins, indicating that salt bridges play significant role in thermostability. Other factors responsible for the protein thermostability include compactness of the protein structure, oligomerization, and strength interaction between subunits [14]. Thermal resistance enzymes determine a free-energy consumption necessary for the transformation from the folded to the unfolded state [15]**.**Thermostability of some enzymes can also be affected by the environmental factors, such as protein concentrations, increased intracellular salts, synthesis of different stabilizers and chemolithoautotrophic mode of nutrition [16].Slight changes in amino acids and sequences increase the stabilizing interactions in the folded protein, such as: additional ion-pairs, disulphide bridges, hydrogen bonds and hydrophobic Interactions [17]. Other stabilizing mechanisms include filing cavities in molecular structure of proteins, shortening of loops reduction of accessible hydrophilic surface area [18]. Thermal resistance of the DNA double helix is greater in hyperthermophiles by reverse gyrase, a unique type I DNA topoisomerase causing positive supertwists for stabilization [19]. In proteases, specific binding of metal ions, particularly calcium, further enhances molecular stability [20]. The abnormally high frequency of tyrosine in thermolysin has also been implicated in its thermostability, although the proposed mechanism appears unique. The stability of thermophilic proteases is not only restricted to temperature but

also includes resistance to denaturing agents, detergents and organic solvents [21]. Other methods for thermophilic proteases through covalent cross linking are well established. Although, adaptation of thermozymes to act at elevated temperatures is mainly achieved by exchange of few amino acid residues and/or their different localization in molecule, the homologous thermostable and thermolabile enzymes are similar and have the same catalytic mechanisms after comparison with the homologous mesophilic proteases [22]. In brief, the main proposed mechanisms/indicators of the increased thermostability include a more highly hydrophobic core, tighter packing or compactness, deleted or shortened loops, greater rigidity, increased Proline content in loops, higher secondary structure content, greater polar surface area, fewer and smaller voids, smaller surface area to volume ratio, fewer thermolabile residues, increased hydrogen bonding, higher isoelectric point, more salt bridges and network of salt bridges [23].More ion pairs have been strongly and consistently linked with thermostability. Water has a dielectric constant of about 80 at 0°C, which drops to 55 at 100°C and is lower still at the extreme pressures near hydrothermal vents in the deep sea where some hyperthermophilic organisms live. A lower dielectric constant makes electrostatic interactions stronger and therefore ion pairs should have a greater stabilizing effect at high temperatures and pressures [24]. Also, better packing has always been a main reason for the higher stability of the thermophilic proteins and hence, smaller and less numerous cavities. One can study the packing of a protein by computing its compactness [25]. Ca^{2+} is a known activator of protease activity and also offers protection against thermal inactivation. All proteases are stabilized by certain levels of free Ca^{+2}. A long-standing goal in Bioinformatics is to identify specific sequences that endow proteases and other proteins with desired functional properties. As opposed to traditional rational and random protein/DNA engineering techniques, many bioinformatic approaches have been developed and used to identify specific changes that influence key functional properties in proteases and many other proteins [26-41]. PROBE and Classifier were used to identify a strand–turn–strand motif consensus sequence within the serine protease subtilase superfamily that appears to endow some thermophilic subtilisins with enhanced thermostability [26].

2. MATERIALS AND METHODS

2.1 Comparative analysis of thermophilic organisms

Retrieval of Organisms

Protease producing thermophilic organisms were retrieved from Genome (https://www.ncbi.nlm.nih.gov/genome/) and Protein (https://www.ncbi.nlm.nih.gov/protein/) databases of NCBI and from UniProt. The NCBI Genome database contains the information on their genomes including sequences, maps, chromosomes, assemblies, and annotations. The NCBI Protein database is a collection of sequences from several sources, including translations from

annotated coding regions in GenBank, RefSeq and Third Party Annotation (TPA), as well as records from SwissProt, PIR, PRF, and PDB [60]. The UniProt database is a freely accessible database of protein sequence and functional information [42].

Classification of Organisms

NCBI Taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi), that provides the information of Organism Classification Ranks [60] was used to classify the organisms based on the taxonomical domain, family and genus. The thermophilic organisms producing proteases were classified into three groups on the basis of temperature: Moderate thermophiles (45-64^{0}C), Extreme thermophiles (65-79^{0}C) and Hyperthermophiles (>80^{0}C) [43].

2.2 Comparative analysis of protein primary sequences

Retrieval of thermophilic proteases sequences

Sequences of Thermophilic proteases were retrieved from NCBI protein database and UniProt Protein primary sequence database in FASTA format. Non-redundant sequences were obtained after removing the redundancy by using CD-Hit (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit).Ninety percent sequence identity cut-off was considered to cluster the sequences. CD-HIT is a very widely used program for clustering and comparing biological sequences [44]. The sequence information such as locus tag and CDs were obtained from NCBI and the Gene name, name of protein, E.C No, sequence length and protease classes were obtained from UniProt.

Retrieval of mesophilic proteases sequences

Sequences of Mesophilic proteases (temperature ranging from 20-40^{0}C) were retrieved from NCBI protein and UniProt database in FASTA format. Non-redundant sequences were obtained after removing the redundancy by using CD-Hit with ninety percent similarity as a cut off. All information (locus tag, CDs, Gene name, name of protein, E.C No, sequence length and protease classes) on the sequence were retrieved from NCBI and UniProt.

Analysis of amino acid frequency between thermophilic and mesophilic proteases

The protease sequences from 74 mesophilic organisms and 74 thermophilic protease producing organisms were compared to understand the amino acid composition and its role in the structural stability of the protein. The sequences of bacteria (38 sequences) and archaea (36 sequences) were analysed separately. The ProtParam (https://web.expasy.org/protparam/) tool was used to predict the amino acid frequency of the proteins. It is a tool which allows the computation of various physical and chemical parameters for a given protein sequence including the amino acid composition [45].

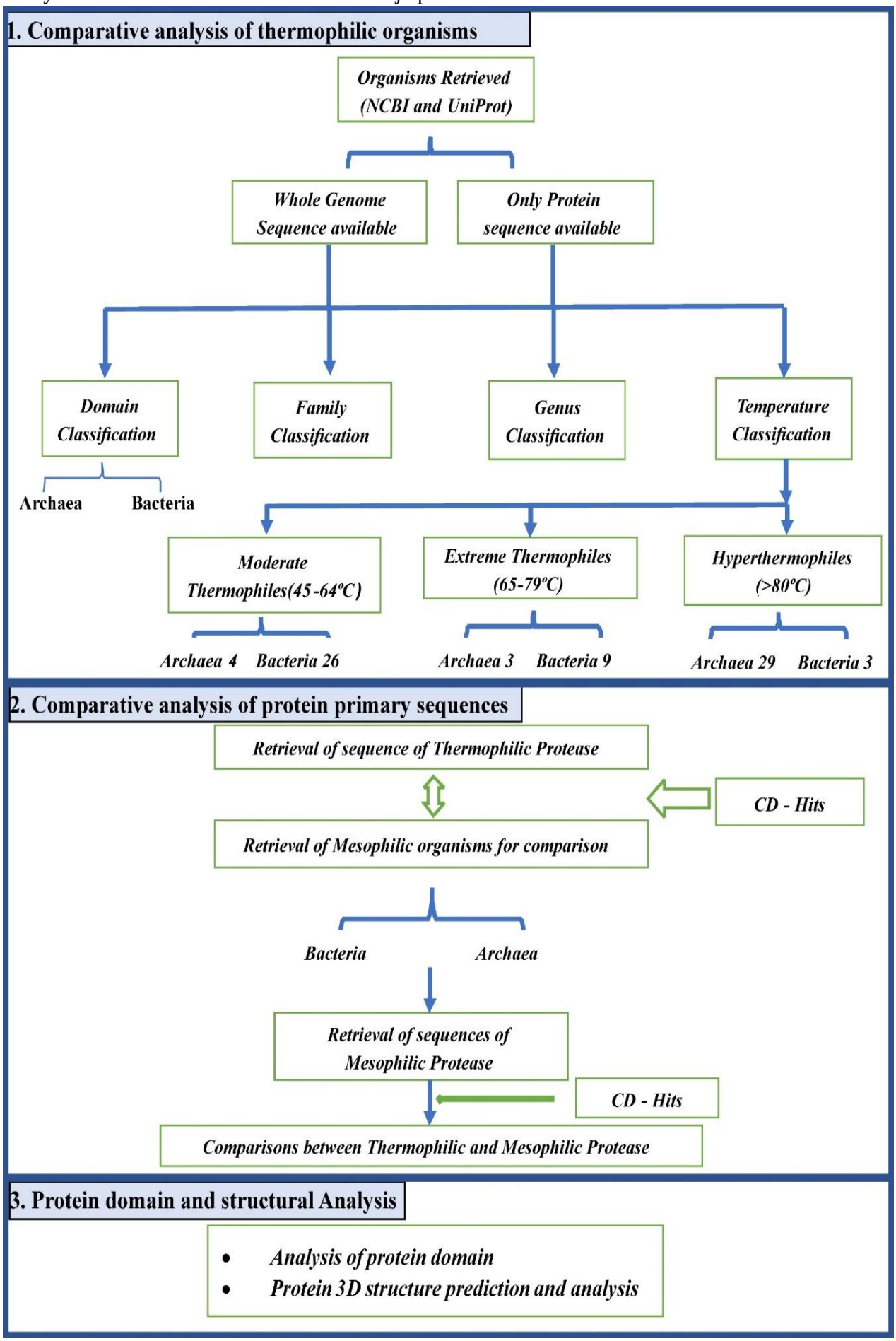

Fig 1: Flow Chart of Materials and methods

2.3 Analysis of protein domain

The protein domains of all sequences were analysed by the InterProScan (https://www.ebi.ac.uk/interpro/search/sequence-search), which provide the insights into the possible functional and essential domains of the proteins [46].

2.4 Protein 3D structure prediction and analysis

As the protein 3D structure of the selected sequences was not available in the Protein Data Bank (PDB), the structures were predicted through homology modelling by SWISS-MODEL (https://swissmodel.expasy.org/interactive). The SWISS-MODEL follows the automated workflows and servers to simplify and streamline the homology modelling process. It generates reliable protein models and have easy access to modelling results, their visualization and interpretation [47]. The quality of the models was evaluated with the Structure Assessment program available at https://swissmodel.expasy.org/assess. This programme evaluates the structure based on the GMQE (Global Model Quality Estimation), QMEAN Z-score, MolProbity and Ramachandran plot. GMQE is a quality estimation which combines properties from the target–template alignment and the template search method. The resulting GMQE score is expressed as a number between 0 and 1, reflecting the expected accuracy of a model built with that alignment and template and the coverage of the target. Higher numbers indicate higher reliability [47]. QMEAN is a composite estimator based on different geometrical properties and provides both global (i.e. for the entire structure) and local (i.e. per residue) absolute quality estimates on the basis of one single model. The QMEAN Z-score provides an estimate of the "degree of nativeness" of the structural features observed in the model on a global scale. It indicates whether the QMEAN score of the model is comparable to what one would expect from experimental structures of similar size. QMEAN Z-scores around zero indicate good agreement between the model structure and experimental structures of similar size. Scores of -4.0 or below is an indication of models with low quality [48]. The MolProbity is a structure-validation web service that provides evaluation of model quality at both the global and local levels for both proteins and nucleic acids. Combined protein quality score that reflects the crystallographic resolution at which such a quality would be expected. The good quality structure has the MolProbity score as low as possible [49, 50]. A Ramachandran plot is a way to visualize energetically favoured regions for backbone dihedral angles against of amino acid residues in protein structure.

3. RESULTS AND DISCUSSION

3.1. Classification of Organisms

Classification based on Domains

Protease producing thermophilic organisms were retrieved from Genome (https://www.ncbi.nlm.nih.gov/genome/) and protein (https://www.ncbi.nlm.nih.gov/protein/) databases of NCBI and UniProt (http://www.uniprot.org/) and were 74 in number. Among these, for 59 organisms, whole genome sequences are available in NCBI (Fig 2). These organisms were

divided into 3 classes on the basis of the temperature range, Moderate thermophiles, Extreme thermophiles and Hyperthermophiles. Moderate thermophilic archaea are 4, while bacteria being 26, Extreme thermophilic archaea 3, and bacteria are 9, Hyperthermophilic archaea are 29 and bacteria are 3. Collectively, the class moderate thermophile, Extreme thermophiles and hyperthermophiles includes 30, 12 and 32 organisms, respectively.

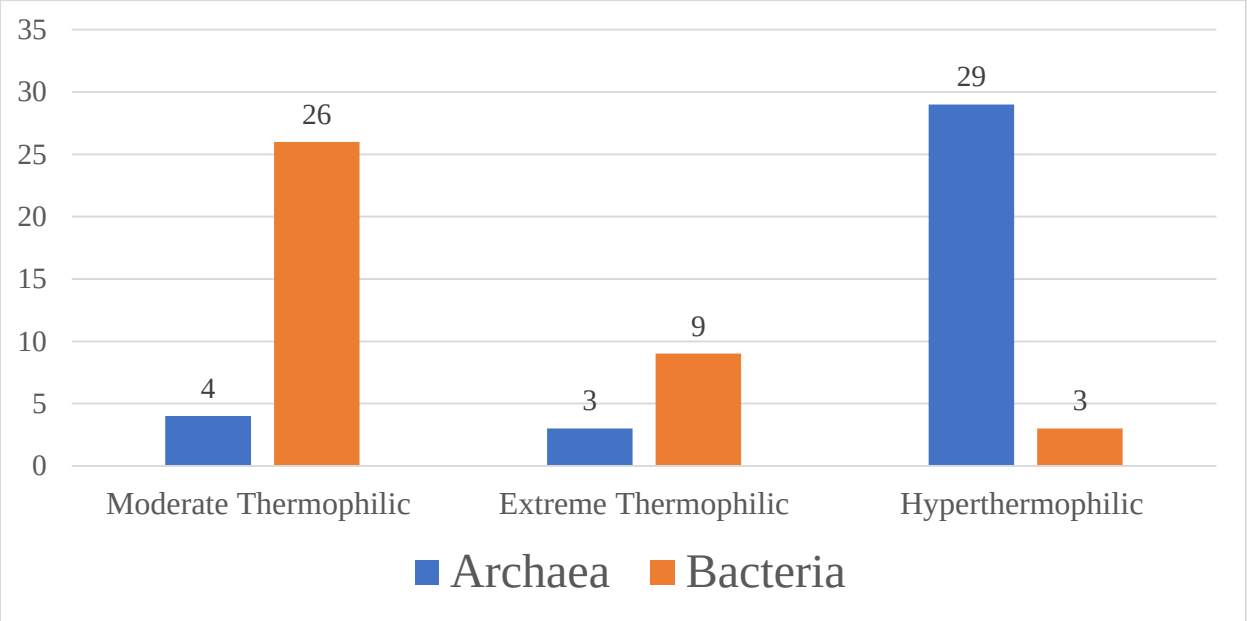

Fig 2: Thermophilic organisms producing proteases are classified on the basis of temperatures into three classes: Moderate thermophiles (Archaea: 4, Bacteria: 26), Extreme thermophiles (Archaea: 3, Bacteria: 9) and Hyperthermophiles (Archaea: 29, Bacteria: 3)

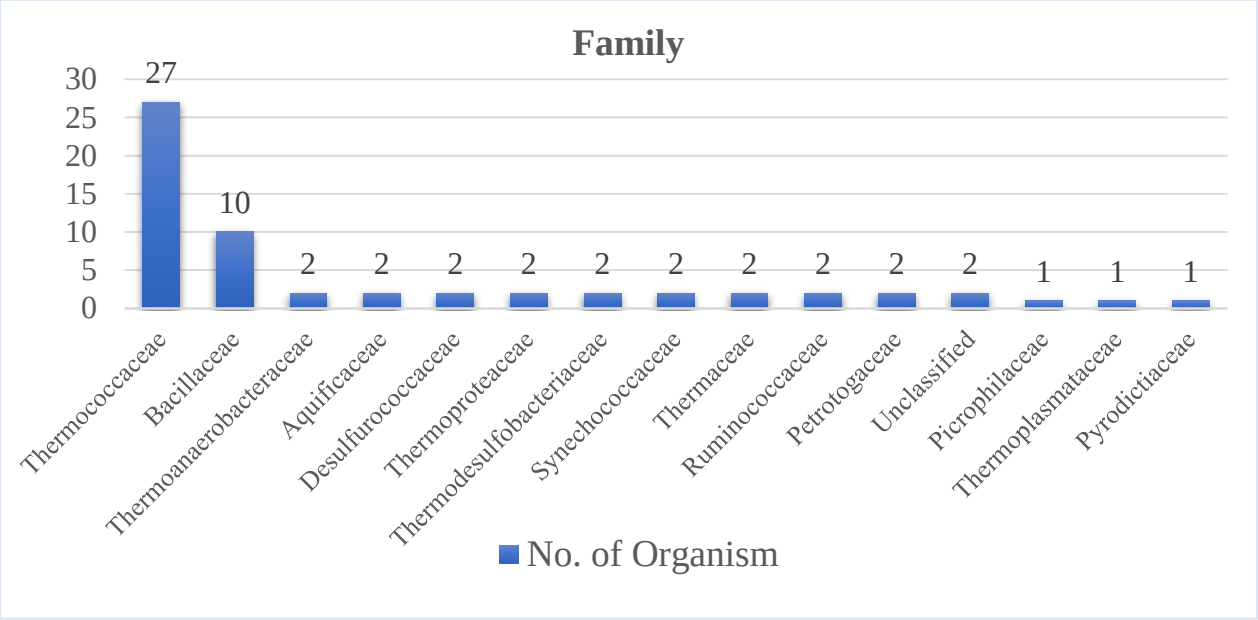

Fig 3a: Classification of thermophiles based on family. All 74 protease producing thermophilic organisms belongs to 29 family. Twenty-seven organisms belong to Thermococcaceae (Archaea)

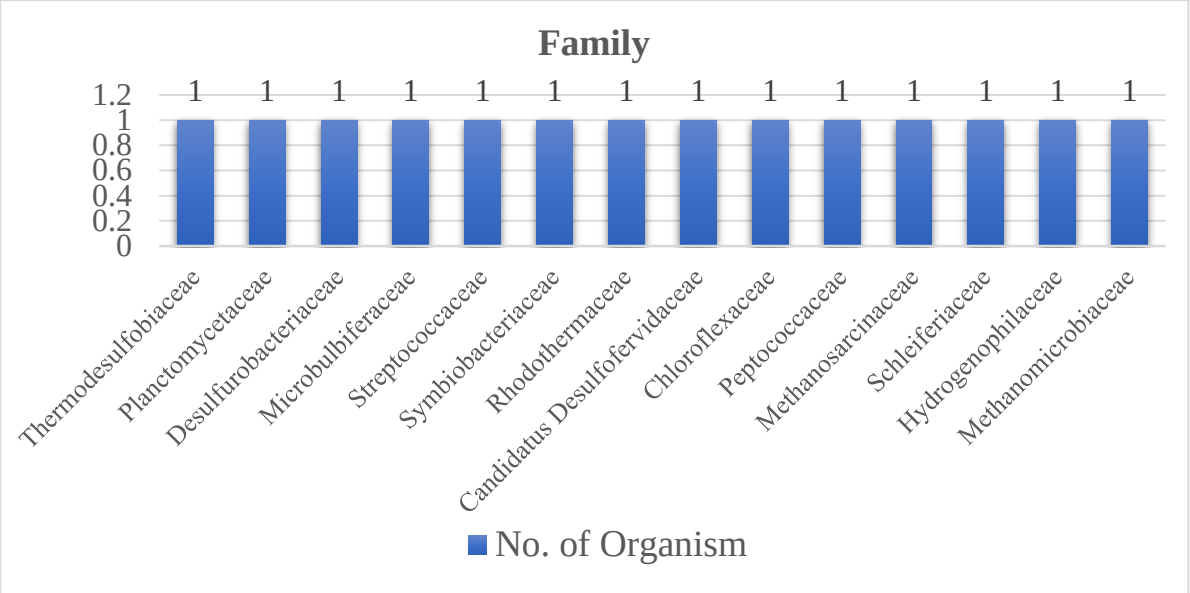

Fig 3b: Continuation of figure 3a

Classification based on Family

The organisms were classified on the basis of their families. The 74 organisms belonged to 29 families among which Thermococcaceae family was associated with the domain archaea, which contains maximum, 27 organisms. Seventeen families had only one organism each (Fig3).

Classification based on Genus

The thermophilic organisms were classified on the basis of the genus. The 74 organisms belonged to 43 genera among which 22 organisms are of Thermococcus genus (Fig 4).

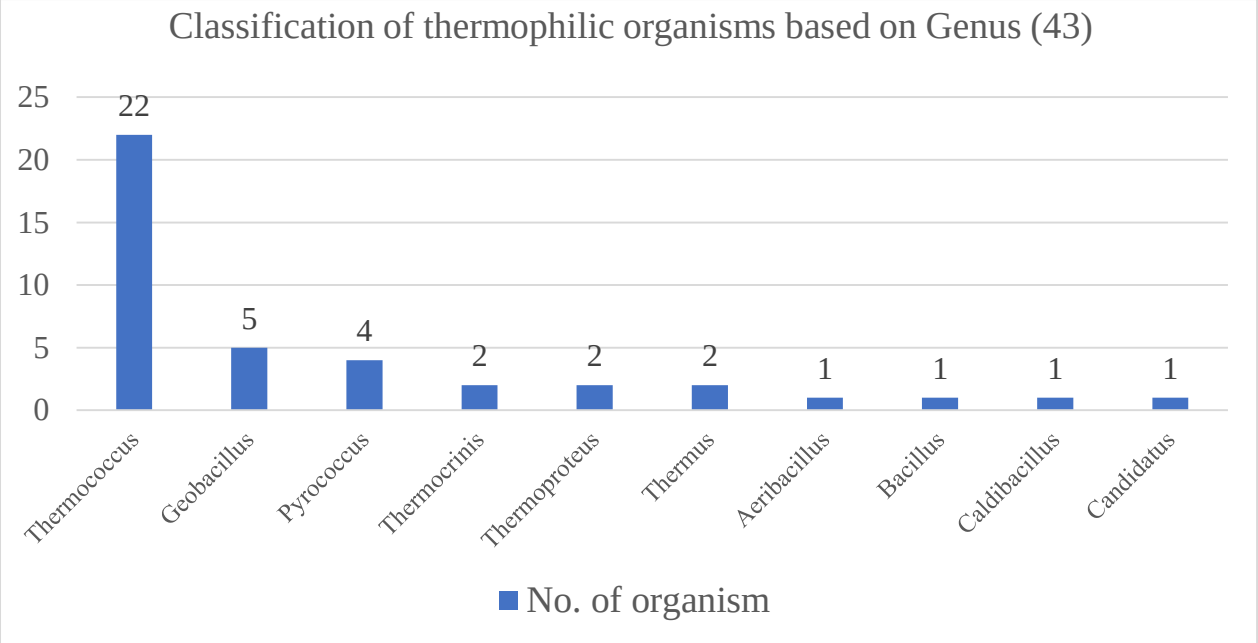

Fig 4a: Classification of the thermophilic organisms producing Protease based on genus. Thermococcus is the genus with maximum number of organisms i.e. 22.

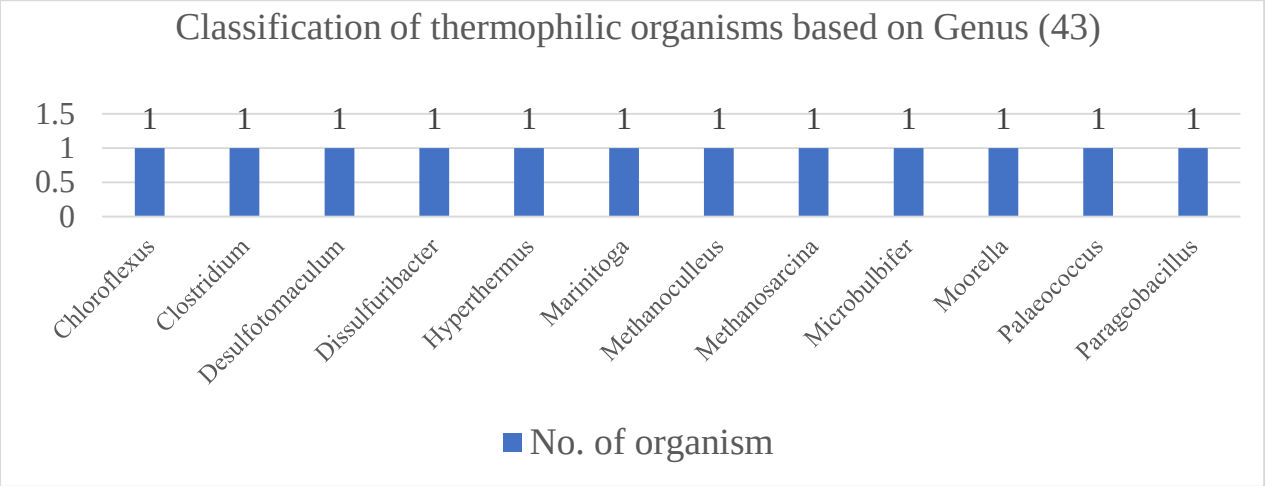

Fig 4b: Continuation of Figure 4a

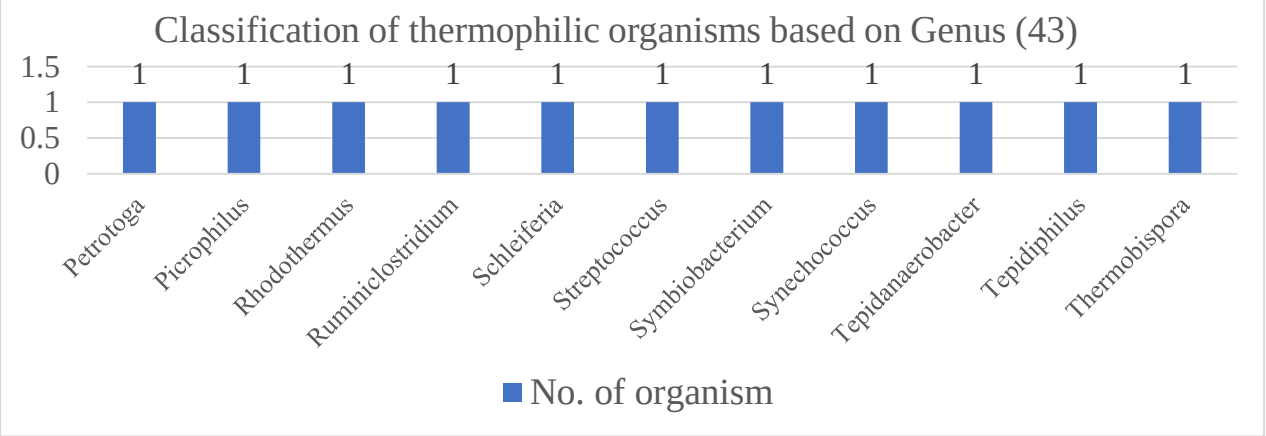

Fig 4c: Continuation of Figure 4b

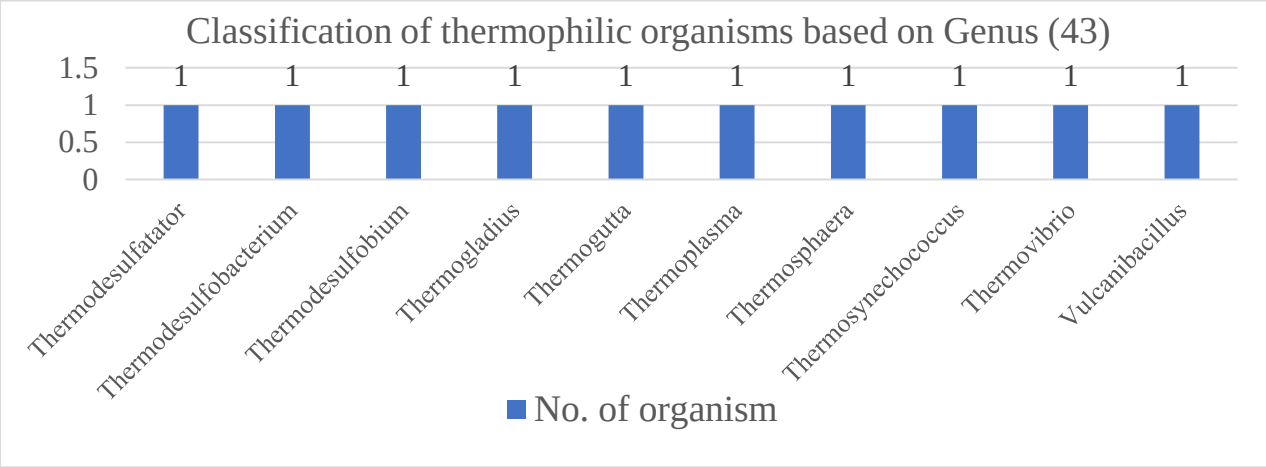

Fig 4d: Continuation of Figure 4c

Classification based on Temperature

The organisms were classified into Moderate Thermophilic Archaea, Moderate Thermophilic Bacteria, Extreme Thermophilic Archaea, Extreme Thermophilic Bacteria, Hyperthermophilic Archaea and Hyperthermophilic Bacteria which included 4, 26, 3, 9, 29 and 3, respectively. To study the stability of the protein produced by thermophiles, they were classified into three classes i.e. Moderate, Extreme and Hyper Thermophiles. The organisms of domain Archaea are maximum

in the class hyper thermophile i.e. the enzyme produced by them would be stable at very high temperatures (>80°C).

3.2. Comparative analysis of protein primary sequences

Primary sequence information of thermophilic organisms

The number of retrieved sequences of protease produced from moderate thermophilic archaea and bacteria, extreme thermophilic archaea and bacteria, hyperthermophilic archaea and bacteria were 19, 68, 4, 16, 38 and 3, respectively. The Sequence length of Thermophilic Proteases retrieved range from 99-1617 AA. The organism and sequence information are tabulated in the table 1, 2 and 3.

Table 1: The Protein sequence information of moderate thermophilic archaea and bacteria

Archaea	**AC_No (UniProt/ NCBI). Number of Amino Acids in brackets**
Methanoculleus thermophilus	A0A1G8ZW87 (793), A0A1G8WR88 (293), A0A1G9BEG1 (662), A0A1G8XBG4 (288), A0A1G8ZFF1 (443), A0A1G8XD83 (303), A0A1G8YIY4 (159), A0A1G8Z6V2 (277), A0A1G9AG36 (351), A0A1G8X031 (171), A0A1G8XFN9 (843)
Methanosarcina thermophila	BAW29844.1 (399), A0A0E3KQ48 (846)
Picrophilus torridus	Q6L2H4 (357)
Thermoplasma acidophilum	P96086 (1071), Q9HJ89 (657), O93655 (780), P96084 (293), Q9HJV2 (317)
Bacteria	
Aeribacillus pallidus	A0A165XNY3 (423)
Bacillus licheniformis	Q65JJ2 (419), Q65H47 (368), Q65I04 (222), Q65EI5 (198), Q65GJ4 (421), Q65KL0 (297), Q65LS8 (206), Q65LS8 (206)
Caldibacillus debilis	A0A1Y3PBN6 (421)
Candidatus Desulfofervidus	A0A127AP35 (357)
Chloroflexus islandicus	A0A178LR89 (201)
Desulfotomaculum ferrireducens	A0A1S6IUD4 (195)
Dissulfuribacter thermophilus	A0A1B9F3B9 (388)
Geobacillus stearothermophilus	WP_095860218.1 (788), P06874 (548), P43133 (551), A0A0K9HTI1 (422)
Geobacillus thermocatenulatus	A0A1V9BJC9 (751)
Geobacillus thermodenitrificans	WP_041264458.1 (788), A4IM88 (180), A4ISQ2 (196)
Marinitoga piezophila	A0A1L7BST0 (405)
Microbulbifer thermotolerans	A0A143HSA0 (497), A0A143HK23 (331)
Moorella thermoacetica	Q2RJP6 (174), Q2RL31 (199), Q2RGQ9 (193), Q2RL30 (419), Q2RKK7 (299)
Petrotoga mobilis	A9BH48 (412), A9BI45 (177)
Ruminiclostridium thermocellum	A3DJR3 (192), A3DJ10 (194), A3DJ11 (431), A3DER9 (99)
Schleiferia thermophila	A0A085L2W8 (681)
Streptococcus thermophilus	Q5M5U4 (196), WP_084828672.1 (428), Q5M5B0 (408), Q5M4Z6 (299), Q5M5L6 (371), A0A0P6V1T9 (380), WP_014727544.1 (1617), WP_014727544.1 (1617)
Symbiobacterium thermophilum	Q67QZ4 (200), Q67SK0 (202), Q67SJ9 (424)
Synechococcus lividus	WP_099798413.1 (199)
Tepidanaerobacter syntrophicus	A0A0U9HRW4 (813)
Tepidiphilus thermophilus	A0A0K6IW40 (808), A0A0K6IMZ8 (178), A0A0K6ISY8 (286), A0A0K6ITQ3 (612), A0A0K6IW01 (208), A0A0K6IXQ1 (102), A0A0K6ITH5 (318), A0A0K6IXW2 (760), A0A0K6IQL1 (219), A0A0K6IPK1 (370)
Thermobispora bispora	D6Y2K4 (927)
Thermodesulfobium narugense	M1E5Z0 (188), M1E9J2 (284)
Thermogutta terrifontis	WP_095415451.1 (193)
Thermosynechococcus elongatus	Q8DHY9 (274), Q8DLI2 (229), Q8DLD3 (205), Q8DJZ9 (198), Q8DLI1 (440)
Vulcanibacillus modesticaldus	A0A1D2YT94 (418)

Table 2: The Protein sequence information of extreme thermophilic archaea and bacteria

Archaea	**AC_No (UniProt). Number of Amino Acids in brackets**
Thermococcus piezophilus	A0A172WJ06 (290)
Thermococcus sibiricus	C6A010 (412)
Thermococcus thioreducens	A0A0Q2REQ8 (290), A0A0Q2QSP6 (286)
Bacteria	
Clostridium stercorarium	A0A1B1YA93 (388)
Geobacillus kaustophilus	Q5L0J5 (421), Q5L0N2 (180), A0A0D8BVC9 (787), Q5KXR9 (225), Q5KVD9 (196), Q5KWJ9 (421)
Geobacillus thermoleovorans	A0A1D7NBU4 (788)
Parageobacillus thermoglucosidasius	A0A1B7KWL8 (422)
Rhodothermus marinus	G2SH40 (199)
Thermocrinis albus	D3SMM9 (289)
Thermodesulfatator indicus	F8ADC1 (356)
Thermus brockianus	A0A1J0LSV5 (309)
Thermus thermophilus	Q72JM6 (804), Q72KS4 (795), Q72L15 (194)

Table 3: The Protein sequence information of hyperthermophilic archaea and bacteria

Archaea	**AC_No (UniProt/NCBI). Number of Amino Acids in brackets**
Hyperthermus butylicus	A2BKW5 (367)
Palaeococcus pacificus	A0A075LUW2 (291)
Pyrococcus abyssi	Q9UZK3 (289), Q9UYC6 (998)
Pyrococcus furiosus	Q8U1S0 (289)
Pyrococcus horikoshii	O58997 (289), O59179 (441), O58221 (1127)
Pyrococcus kukulkanii	A0A127BB69 (264)
Thermococcus barophilus	A0A0S1X9V0 (292)
Thermococcus barossii	WP_088864647.1 (290), WP_088864646.1 (301), WP_088864504.1 (285)
Thermococcus celer	A0A218P1F0 (289)
Thermococcus chitonophagus	A0A161KAG4 (289)
Thermococcus cleftensis	I3ZWV1 (290)
Thermococcus eurythermalis	A0A097QUW8 (299)
Thermococcus gammatolerans	C5A5K3 (291)
Thermococcus gorgonarius	WP_088885308.1 (290)
Thermococcus guaymasensis	A0A0X1KM06 (291)
Thermococcus kodakarensis	Q5JEZ8 (290)
Thermococcus litoralis	H3ZKT8 (292)
Thermococcus nautili	W8NT40 (291)
Thermococcus onnurineus	B6YVB5 (290)
Thermococcus pacificus	A0A218P8U4 (290)
Thermococcus paralvinellae	W0I6H0 (292)
Thermococcus peptonophilus	A0A142CT47 (290)
Thermococcus profundus	WP_088858287.1 (290)
Thermococcus radiotolerans	WP_088866347.1 (286), WP_088866211.1 (301), WP_088866210.1 (292)
Thermococcus siculi	WP_088856898.1 (444), WP_088856465.1 (290), WP_088856464.1 (301)
Thermogladius calderae	I3TGE5 (385)
Thermoproteus uzoniensis	F2L4V0 (346)
Thermoproteus tenax	G4RM24 (308)
Thermosphaera aggregans	D5U048 (398)
Bacteria	
Thermocrinis ruber	W0DB03 (288)
Thermodesulfobacterium geofontis	F8C1T2 (286)
Thermovibrio ammonificans	E8T5U6 (426)

Primary sequence information of mesophilic organisms

A total of 74 sequences of Protease produced from Mesophilic organisms were obtained. Among them, 38 were Bacteria and 36 Archaea. Seventy-Two protein sequences along with their information were collected. The sequence length of the Mesophilic Proteases retrieved range from 98 – 1088. The organism and sequence information are tabulated in the table 4 and 5.

Table 4: The Protein sequence information of mesophilic archaea

Thermotolerant/Mesophilic (20-40°C) Archaea	Ac_No (UniProt / NCBI). Number of Amino Acids in brackets
Halostagnicola kamekurae	A0A1I6SAW3 (274)
Halogranum amylolyticum	WP_089824074.1 (197), A0A1H8V7B1 (276)
Halopelagius inordinatus	A0A1I2P3F0 (196)
Halopelagius longus	A0A1H1B8A8 (196)
Natronoarchaeum philippinense	WP_097008881.1 (274), A0A285P7Z2 (197)
Halomicrobium zhouii	WP_089813170.1 (274), A0A1I6LSP9 (196)
Halorubrum chaoviator	A0A238Z3A0 (293)
Halobellus clavatus	WP_089768423.1 (196), A0A1H3EP03 (274)
Halostagnicola larsenii	W0JJY0 (274)
Halovenus aranensis	WP_092701941.1 (196), A0A1G8Y0V0 (284)
Natronorubrum texcoconense	WP_090305202.1 (274), A0A1G9E9M3 (198)
Natronorubrum sediminis	A0A1H6FNT8 (274)
Haloplanus vescus	WP_092632809.1 (273)
Haloarcula hispanica	V5TMH6 (381), A0A165L7Y1 (274), G0HVF0 (354)
Pyrodictium occultum	A0A0V8RVV3 (366), A0A0V8RRH9 (274)
Natronorubrum thiooxidans	A0A1N7EIS1 (199), A0A1N7F691 (274)
Natrinema salaciae	WP_090613786.1 (197), A0A1H9QH96 (274)
Halorubrum tropicale	A0A0M9AS85 (289), 0A0N0BR31 (758), A0A0M9AIJ2 (293), A0A0M9AQH4 (293), A0A0M9ARX1 (208), A0A0M9APH0 (224)
Haloarchaeobius iranensis	A0A1G9YGY1 (276)
Haloarcula rubripromontorii	A0A0N1IUH9 (377), WP_053968755.1 (274), A0A0M9ALB1 (353)
Halorubrum sodomense	A0A1I6FLG5 (197)
Halorubrum vacuolatum	WP_089383917.1 (216), A0A238X4B0 (265)
Halorientalis regularis	WP_092687056.1 (198), A0A1G7NTH0 (272)
Haladaptatus litoreus	A0A1N6VHJ7 (177), A0A1N6XUI8 (197)
Methanosarcina barkeri	Q469F5 (802)
Haloferax volcanii	D4GXB9 (198)
Haloferax mucosum	WP_008320990.1 (198)
Haloarcula vallismortis	A0A1H2XTN4 (196), M0JRE6 (273)
Haloarcula californiae	M0KUS5 (196), WP_049944562.1 (352)
Haloferax mediterranei	WP_004057186.1 (198), I3R794 (519)
Natrinema pellirubrum	L0JP36 (274), L0JL06 (197)
Haloarcula marismortui	Q5V2G7 (381), Q5V665 (274)
Candidatus Methanomethylophilus	A0A0W7TKD2 (183), M9SJE6 (634), A0A0W7TI19 (317), A0A0W7TKD2 (183)
Halobiforma haloterrestris	WP_089784801.1 (274), A0A1I1IIG6 (197)
Methanosarcina mazei	WP_015412027.1 (868), Q8PSE5 (294), Q8PXI2 (287)
Sulfolobus solfataricus	P95871 (1068), Q97UA2 (310), Q97X95 (311), Q97TZ9 (325)

Table 5: The Protein sequence information of mesophilic bacteria

Thermotolerant/Mesophilic (20-40°C) Bacteria	**Ac_No (UniProt / NCBI). Number of Amino Acids in brackets**
Acinetobacter baylyi	WP_088459338.1 (192), WP_004922314.1 (192), Q6FAX7 (383), Q6F8Q1 (301), Q6FEP8 (201), Q6FEP7 (436), Q6FCI2 (107)
Bacillus altitudinis	A0A0J1I7T1 (422)
Rhodobacter maris	WP_097070245.1 (196), A0A285T471 (185)
Bacillus thermozeamaize	A0A1Y3PMR1 (188), A0A1Y3PT95 (368), A0A1Y3PRH9 (304), A0A1Y3PEB0 (422), A0A1Y3PLY4 (197), A0A1Y3PE82 (198), A0A1Y3PVA2 (814), A0A1Y3PQG4 (416), A0A1Y3PQJ9 (430), A0A1Y3PNQ3 (170), A0A1Y3PJM3 (222), A0A1Y3PPQ1 (230)
Bathymodiolus thermophilus	A0A1J5U8G0 (778), A0A1J5TV25 (427), A0A1J5TWT5 (198), A0A1J5TY82 (98), A0A1J5UA42 (207), A0A1J5UCX2 (739)
Bellilinea caldifistulae	A0A0N8GNI0 (364), A0A0P6XJ65 (811), A0A0P6WZZ2 (809), A0A0P6WU62 (201), A0A0N8GNI5 (179), A0A0P6X2A6 (818), A0A0N8GNC8 (685), A0A0P6X2Y0 (315)
Brevibacillus borstelensis	WP_031931719.1 (412), M8DD12 (412)
Brochothrix thermosphacta	A0A1D2JRB3 (411), A0A1D2LQA7 (411)
Candidatus Kryptobacter	A0A0N7MWE0 (1088)
Fluoribacter gormanii	A0A0W0U2B8 (598)
Mycobacterium goodii	A0A0K0X9P0 (778), A0A0K0X7L5 (426), A0A0K0X7D1 (219), A0A0K0X6T1 (100), A0A0K0X9K1 (318), A0A0K0XE11 (434), A0A0K0X118 (396), A0A0K0X4U3 (727), A0A0K0X6K0 (498), A0A0K0XGX9 (449), A0A0K0X1G5 (260), A0A0K0XF80 (340), A0A0K0XGX3 (521), A0A0K0X9Y8 (383), A0A0K0X030 (208), A0A0K0X9C7 (188)
Mycobacterium thermoresistibile	A0A100XBS2 (523)
Paenibacillus naphthalenovorans	A0A0U2MY05 (427)
Streptomyces hygroscopicus	H2JYN1 (329), A0A0S2P1Z4 (329)
Tepidibacillus decaturensis	A0A135L2D7 (421)
Brevundimonas bacteroides	WP_029417646.1 (336)
Bradyrhizobium japonicum	WP_081369219.1 (874), WP_018643188.1 (857)
Paraburkholderia nodosa	WP_069267890.1 (193)
Paraburkholderia phenazinium	A0A1N6H7Y0 (193)
Achromobacter denitrificans	A0A1Z3H764 (192)
Achromobacter piechaudii	D4XHL4 (192)
Burkholderia pyrrocinia	WP_017334196.1 (193)
Listeria rocourtiae	WP_077913470.1 (443)
Burkholderia ubonensis	WP_095403930.1 (192), WP_095400577.1 (193)
Enterococcus faecalis	WP_048949418.1 (415), WP_002414438.1 (444), Q834K3 (182), Q834K4 (467), Q837R0 (197), Q833M7 (417)
Burkholderia multivorans	WP_069219930.1 (895), B9BTW4 (193), A9AC32 (178), A9AJR1 (423), A9AC67 (285), A9AGM9 (104)
Lactobacillus acidophilus	Q5FL55 (195), Q5FMS7 (298)
Stappia indica	WP_088946061.1 (182), A0A285RA60 (189)
Paraburkholderia diazotrophica	A0A1H6QE57 (193)
Mesorhizobium qingshengii	WP_091582710.1 (197), A0A1G5ZEV2 (206)
Paracoccus tibetensis	A0A1G5HGH4 (197)
Noviherbaspirillum humi	WP_089397867.1 (457)
Loktanella litorea	A0A1I6M744 (216)
Enterococcus mundtii	A0A1L8UQE6 (413)
Caballeronia sordidicola	WP_089164707.1 (193)
Sporomusa acidovorans	WP_093797061.1 (189)
Hoeflea halophila	A0A286IAV1 (653)
Desulforhopalus singaporensis	A0A1H0KNW5 (195)

Analysis of amino acid frequency between thermophilic and mesophilic proteases

Comparison of Thermophilic and Mesophilic sequences is an integral step of analysing the stability of different amino acids in both organisms. It was necessary to consider 74 Mesophilic Protease producing organisms as an alternate to 74 Thermophilic Protease producing organisms. Among them, 38 were Bacteria and 36 Archaea. Sequences with more or less same length of amino acids from Thermophilic and Mesophilic organisms were considered for the average comparisons between the two. The Amino Acids were classified on the basis of eight groups (Table 7). There are no specific rules of amino acid composition for the stability of the protein but on the basis of many experimental evidences as a result of the experiments carried between hyper thermophilic and mesophilic protein, it is estimated that Aliphatic amino acids, such as A and L are normally found less while G, V, and I are found more in Thermostable protein. Amino Acid composition studies were compared with other Literature studies (Table 6).

Table 6: Amino Acid Composition based on Literature Studies (9, 22)

Name of the amino acids	Three letter Symbol	One letter Symbol	Thermophiles	Mesophiles
Glycine	Gly	G	More	Less
Alanine	Ala	A	Less	More
Valine	Val	V	More	Less
Isoleucine	Ile	I	More	Less
Leucine	Leu	L	Less	More
Methionine	Met	M	More	Less
Proline	Pro	P	More	Less
Tryptophan	Trp	W	Less	More
Phenylalanine	Phe	F	Less	More
Tyrosine	Tyr	Y	More	Less
Arginine	Arg	R	More	Less
Lysine	Lys	K	More	Less
Histidine	His	H	Less	More
Glutamic Acid	Glu	E	More	Less
Aspartic Acid	Asp	D	Less	More
Serine	Ser	S	Less	More
Threonine	Thr	T	Less	More
Cysteine	Cys	C	Less	More
Glutamine	Gln	Q	Less	More
Asparagine	Asn	N	Less	More

Analysing the results of the amino acid frequency between mesophilic and thermophilic archaea on the basis of the literature, it is found that there are some amino acids whose frequency contradict the literature values. As per the literature, the frequency of D should be less in thermophiles, while the results in this report show greater values in moderate thermophile. Similarly, M, N and F, as suggested in literature, is found less frequently while the results in this report suggest higher frequency than that of mesophilic protein. According to literature, the amino acids R and P are higher in moderate thermophile, but in this case, it's at lower level. The comparison between the mesophilic and thermophilic archaea shown in figure 11.

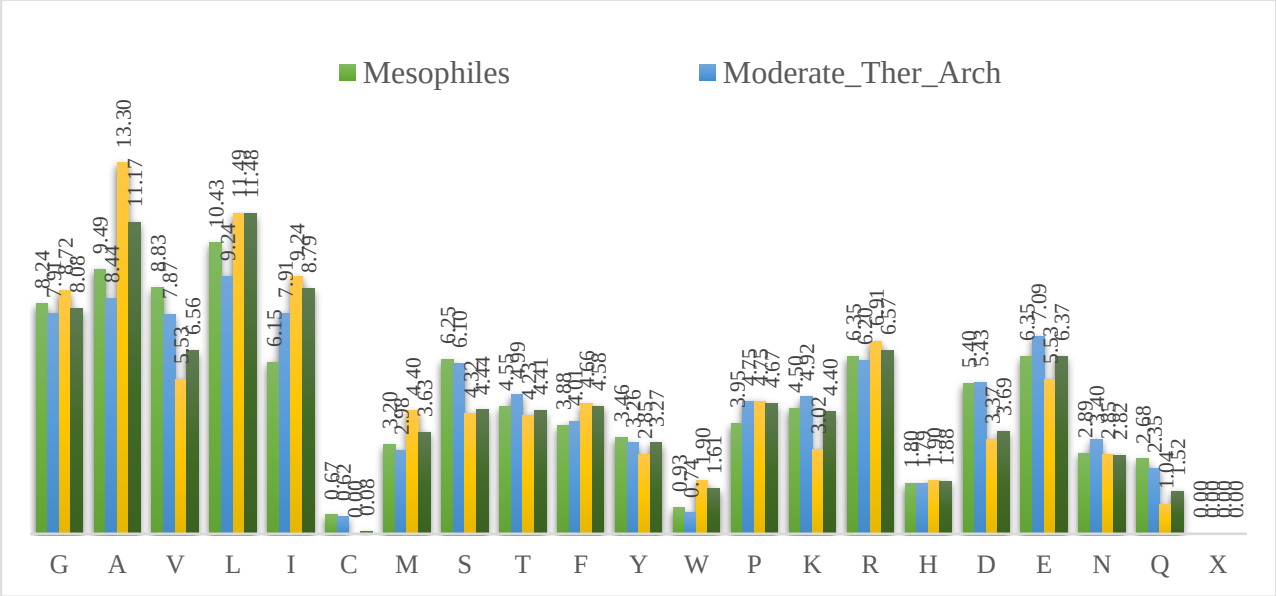

Fig 11: Comparison of amino acid frequency between Mesophilic Archaea – Moderate Thermophilic Archaea – Extreme Thermophilic Archaea – Hyperthermophilic Archaea

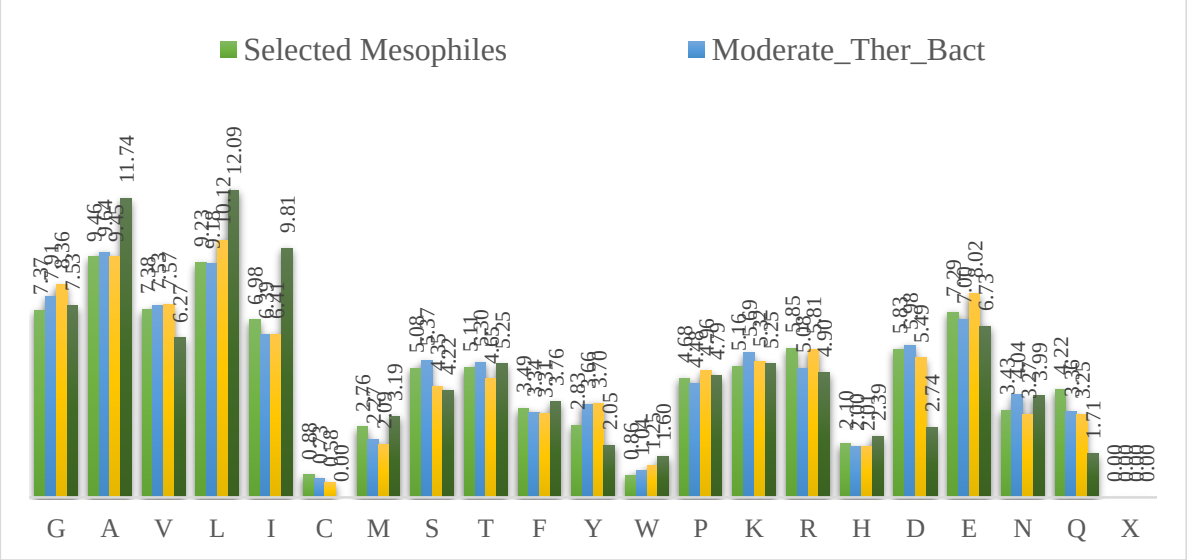

Fig 12: Comparisons of amino acid frequency between Mesophilic Bacteria – Moderate Thermophilic Bacteria – Extreme Thermophilic Bacteria – Hyperthermophilic Bacteria

Analysing the results of the amino acid frequency between mesophilic and thermophilic bacteria, its reflected that the amino acids: G, C and K correspond to the similar level as in literature, while A,

W and Q reflect an opposite trend. V and H display different results for Hyperthermophilic bacteria, L shows opposite results, while I show difference in Moderate and Extremethermophilic bacteria. The amino acid M in moderate and extreme thermophile shows different pattern, a greater number of S and D present in moderate thermophilic bacteria. The hyperthermophilic bacteria show F and Y different pattern, and moderate thermophilic bacteria show the amino acid P at lower level as shown in figure 12. The Comparisons of amino acid frequency between Mesophilic Archaea-Moderate thermophilic Archaea and Mesophilic bacteria –Moderate thermophilic bacteria revealed that the frequency of Q and H in bacteria is as per the literature, i.e. it must be less in thermophilic rather than mesophilic while for archaea it is opposite to literature. In Archaea the amino acid M is different, and the A, L were opposite in Bacteria. In Archaea, the frequency of D and S is as per the literature, i.e. it must be less in thermophilic rather than mesophilic while for bacteria it is opposite to the literature, and T was found opposite. As per literature studies, the frequency of amino acid G should be more in thermophilic than in mesophilic proteins, a pattern which is reflected in bacteria but not for archaea. The frequency of amino acid R and E in Bacteria is different than the literature as shown in figure 13. The comparison of the amino acid frequency between mesophilic-extreme thermophilic archaea and mesophilic-extreme thermophilic bacteria revealed that the amino acids G, C, S, Y, W, P, H, D, Q are in greater number in thermophilic than that of mesophilic as. The amino acids A, L, T, F, K, R, N are opposite, the amino acid V is less frequent in extreme thermophilic archaea (Fig 14).

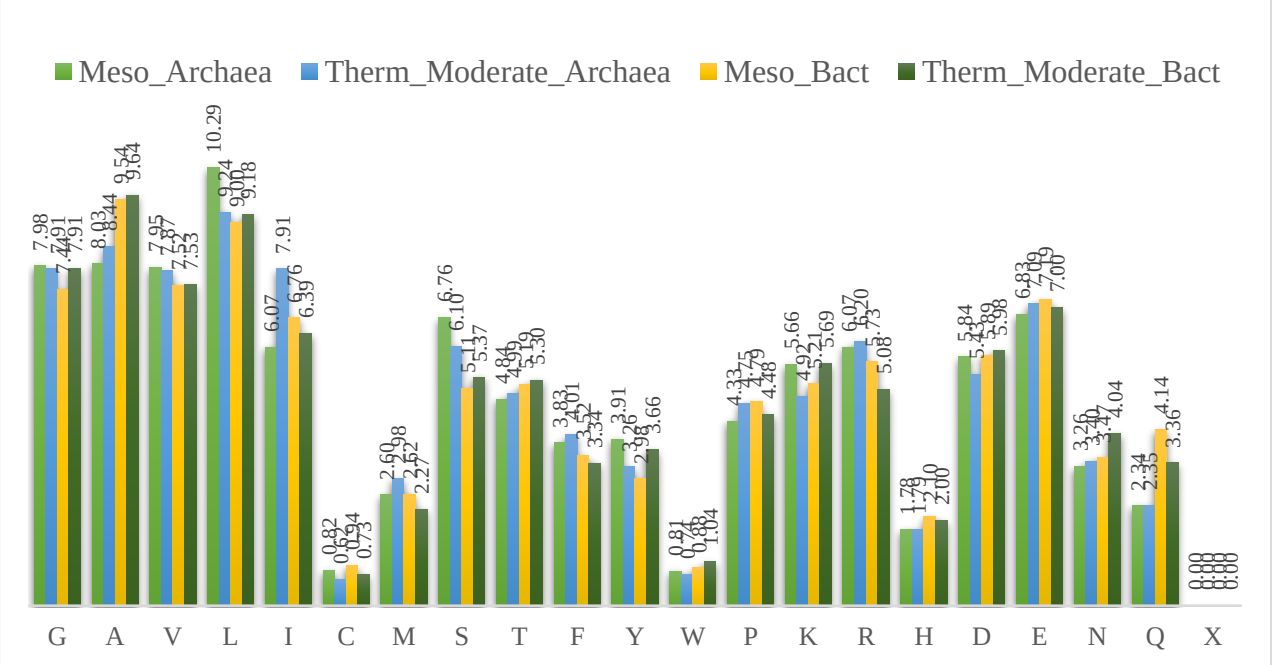

Fig 13: Comparisons of amino acid frequency between Mesophilic Archaea-Moderate thermophilic Archaea and Mesophilic bacteria –Moderate thermophilic bacteria

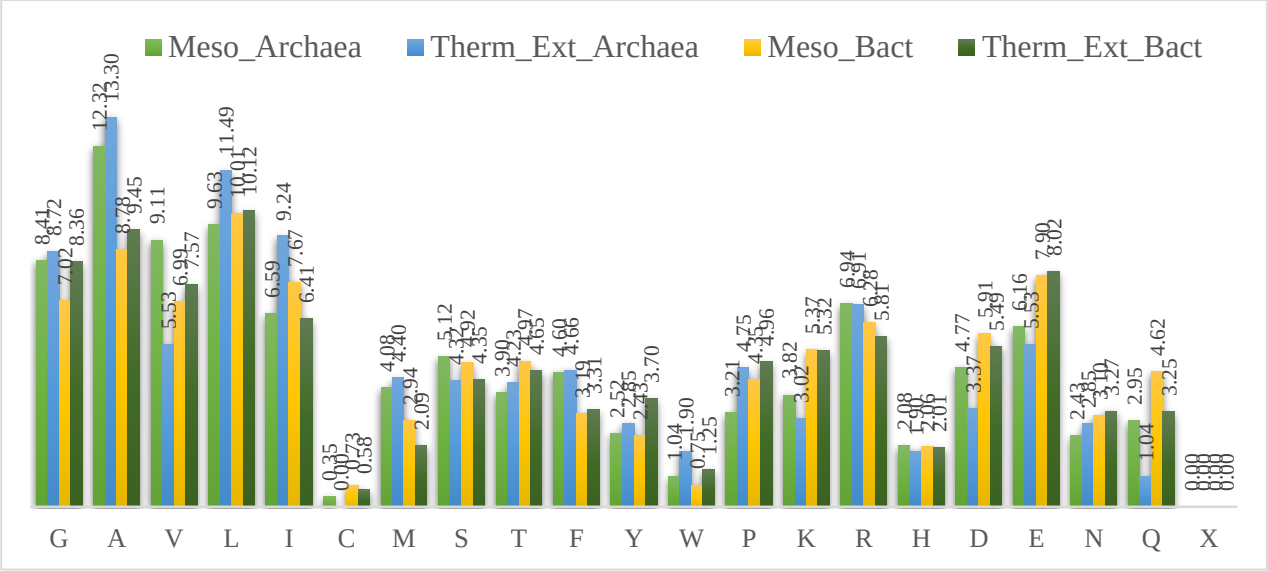

Fig 14: Comparisons of amino acid frequency between Mesophilic Archaea-Extreme thermophilic Archaea and Mesophilic Bacteria – Extreme thermophilic Bacteria

Comparisons of amino acid frequency between Mesophilic Archaea - Hyperthermophilic Archaea and Mesophilic Bacteria – Hyperthermophilic shows that the amino acids G, V, L, T, H, N are opposite in the results than shown in the literature. The amino acids A and F are different in archaea and the I, C, S, Y, W, P, K, D, E, Q are same as per the literature, while M is different in case of Archaea as shown in figure 15.

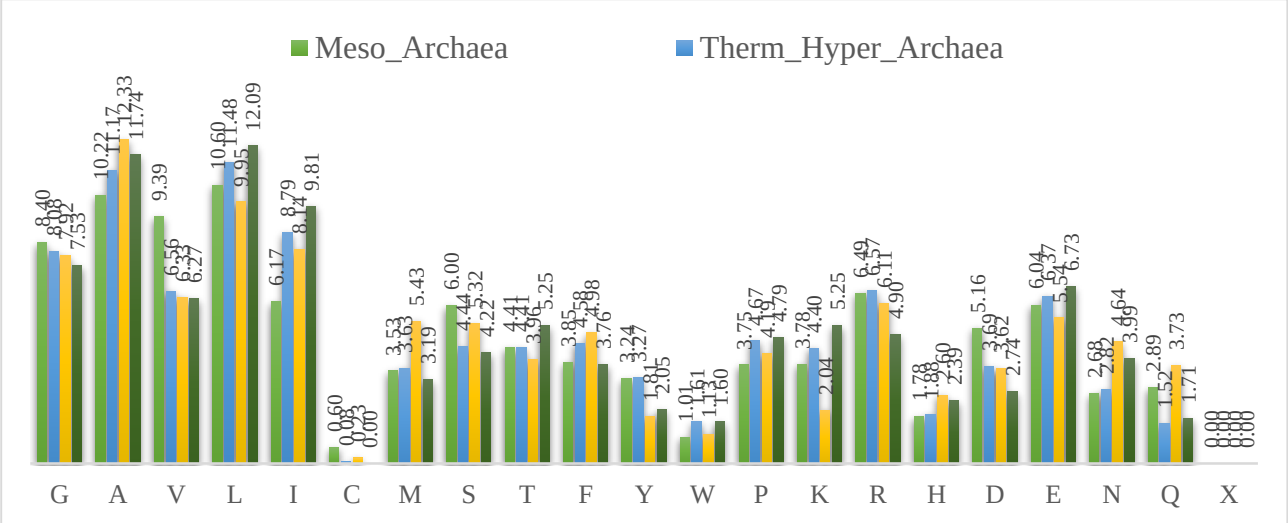

Fig 15: Comparisons of amino acid frequency between Mesophilic Archaea - Hyperthermophilic Archaea and Mesophilic Bacteria – Hyperthermophilic Bacteria

Table 7: Classification of Amino Acids on the basis of different groups

Aliphatic	G A V L I
Sulphur Containing	C M
Hydroxy	S T
Aromatic	F Y W
Heterocyclic	P
Basic	K R H
Acidic	D E N Q
Unclassified	X

The results obtained after the comparative studies between Thermophilic and Mesophilic counterparts show that Alanine and Leucine are relatively found more in Thermophiles which is against the theory of Literature evidences although literature doesn't specify the protein on which their studies have been done. The non-polar amino acids are characterized by having no polar atoms (only carbon and hydrogen) in their side chains. They include Glycine (Gly, G), Ala (Alanine, A), Val (Valine, V), Leu (Leucine, L), Ile (Isoleucine, I), Pro (Proline, P), and Met (Methionine, M). Among hydrophobic residues, Ala, Val, Leu, and Ile belong to the aliphatic amino acids. It has been widely accepted that the aliphatic amino acids would contribute to the hydrophobic interaction, which is the main force for maintaining conformational stability in the inner part of protein. Some analysis showed that thermophilic proteins have a higher frequency of proline. The proline has been used to increase the protein stability in the several mutational studies and hence an increase of Pro content may be due to the increase of the thermophilic protein rigidity. This indicated that the rigid conformation or the turn conformation of thermophilic proteins might have better packed forms than those of mesophilic ones. The Met is known as thermolabile amino acid due to its tendency to undergo oxidation at high temperature. The polar nature of the side chain means that these amino acids are ready to interact with water (hydrophilic). This group includes Asparagine (Asn, N), Glutamate (Gln, Q), Cysteine (Cys, C), Serine (Ser, S) and Threonine (Thr, T). Asn and Gln are known as thermolabile amino acids due to their tendency to undergo deamination at high temperature. Ser and Thr are known as the best residues for interacting with the waters surrounding protein. The side group of Cys also contains a sulfur atom. But thermophilic proteins showed a higher frequency of Glu both in buried and exposed state. Related with another higher frequency of Arg, the higher frequency of Glu could be explained as counter trend for making a salt-bridge in thermophilic proteins.

3.3.Analysis of protein domain

The InterProScan revealed various domains of the obtained proteins. As a case study, only one sequence in every category as shown in table 8 was undertaken. The Peptidase M48 domain predicted in the hyperthermophilic archaea, represents the largely extracellular catalytic region of CAAX prenyl protease homologues such as Human FACE-1 protease. These are metallopeptidases, with the characteristic HExxH motif giving the two histidine-zinc-ligands and an adjacent glutamate on the next helix being the third. The whole molecule folds to form a deep groove/cleft into which the substrate can fit. This group of metallopeptidases belongs to MEROPS peptidase family M48. Proteins with this domain are mostly described as a probable protease htpX homologue (EC:3.4.24) or CAAX prenyl protease 1, which proteolytically removes the C-terminal three residues of farnesylated proteins. They are integral membrane proteins associated with the endoplasmic reticulum and Golgi, binding one zinc ion per subunit [51, 52]. The peptidase M50 domain was

mostly observed in all sequences except hyperthermophilic archaea. The enzymes having peptidase M50 domain, a divalent cation which is usually zinc, but may be cobalt, manganese or copper, activates the water molecule. The metal ion is held in place by amino acid ligands, usually three in number. In some families of co-catalytic metallopeptidases, two metal ions are observed in crystal structures ligated by five amino acids, with one amino acid ligating both metal ions. The known metal ligands are His, Glu, Asp or Lys. At least one other residue is required for catalysis, which may play an electrophillic role. Many metalloproteases contain an HEXXH motif, which has been shown in crystallographic studies to form part of the metal-binding site. The HEXXH motif is relatively common, but can be more stringently defined for metalloproteases as 'abXHEbbHbc', where 'a' is most often valine or threonine and forms part of the S1' subsite in thermolysin and neprilysin, 'b' is an uncharged residue, and 'c' a hydrophobic residue. Proline is never found in this site, possibly because it would break the helical structure adopted by this motif in metalloproteases [53]. The PDZ domain (also known as Discs-large homologous regions (DHR) or GLGF) found in all bacteria except mesophile. PDZ domains are found in diverse signalling proteins in bacteria, yeasts, plants, insects and vertebrates. PDZ domains can occur in one or multiple copies and are nearly always found in cytoplasmic proteins. They bind either the carboxyl-terminal sequences of proteins or internal peptide sequences. In most cases, interaction between a PDZ domain and its target is constitutive, with a binding affinity of 1 to 10 microns. However, agonist-dependent activation of cell surface receptors is sometimes required to promote interaction with a PDZ protein. PDZ domain proteins are frequently associated with the plasma membrane, a compartment where high concentrations of phosphatidylinositol 4,5-bisphosphate (PIP2) are found. Direct interaction between PIP2 and a subset of class II PDZ domains (syntenin, CASK, Tiam-1) has been demonstrated. PDZ domains consist of 80 to 90 amino acids comprising six beta-strands (beta-A to beta-F) and two alpha-helices, A and B, compactly arranged in a globular structure. Peptide binding of the ligand takes place in an elongated surface groove as an anti-parallel beta-strand interacts with the beta-B strand and the B helix. The structure of PDZ domains allows binding to a free carboxylate group at the end of a peptide through a carboxylate-binding loop between the beta-A and beta-B strands [54, 55]. The CBS domain was observed in mesophilic bacteria.CBS domains are small intracellular modules that pair together to form a stable globular domain. Pairs of these domains have been termed a Bateman domain. CBS domains have been shown to bind ligands with an adenosyl group such as AMP, ATP and S-AdoMet. CBS domains are found attached to a wide range of other protein domains suggesting that CBS domains may play a regulatory role making proteins sensitive to adenosyl carrying ligands. The region containing the CBS domains in cystathionine-beta synthase is involved in regulation by S-AdoMet. CBS domain pairs from AMPK bind AMP or ATP. The CBS domains from IMPDH and the chloride channel CLC2 bind ATP [56-59].

3.4. Protein 3D structure prediction and analysis

All selected proteins were homology modelled by SWISS-MODEL and evaluated the quality of predicted structure through the structural assessment server of swiss model workspace. The details of the structure and the evaluation data are given in the table 9. The predicted structure shows all the archaeal protease required metal ion Zn for their optimum activity and the bacterial protease may not require the metal ion Zn for its function except mesophilic bacteria. The three amino acids Glu, Asp (Acidic) and His (Basic) are involved in the binding with Zn metal ion. The metal ion significantly helps the stability and activity of the protease at high temperatures. The score of GMQE, QMEAN, MolProbity, Ramachandran Favoured shows the quality of predicted structures are significantly better.

Table 8: Protein domain predicted by InterProScan and transmembrane region extracted from UniProt

Class	Group	Organism	Family	Temperature (°C)	Opt. Temp (°C)	AC_No (UniProt / NCBI)	No. of AA	Domain Analysis (InterProScan)	Transmembrane region
Hyperthermophiles (>80°C)	Archaea	Pyrococcus kukulkanii	Thermococcaceae	70-112	105	A0A127BB69	264	Peptidase M48 (IPR001915) (AAs: 58 - 253)	6 – 34 (29 aa), 157 – 174 (18 aa)
Hyperthermophiles (>80°C)	Bacteria	Thermovibrio ammonificans	Desulfurobacteriaceae	60-80	85	E8T5U6	426	Peptidase M50 (IPR008915) (AAs: 5 - 205 & 271 - 412), PDZ domain (IPR001478) (AAs: 111 -270)	6 – 25 (20 aa), 90 – 114 (25 aa), 364 – 385 (22 aa), 406 – 425 (20 aa)
Extreme Thermophiles (65-79°C)	Archaea	Thermococcus sibiricus	Thermococcaceae	40-88	78	C6A010	412	Peptidase M50 (IPR008915) (AAs: 300 - 336)	129 – 148 (20 aa), 168 – 185 (18 aa), 231 – 253 (23 aa), 265 – 284 (20 aa), 296 – 320 (25 aa), 340 – 368 (29 aa), 389 – 410 (22 aa)
Extreme Thermophiles (65-79°C)	Bacteria	Thermodesulfatator indicus	Thermodesulfobacteriaceae	55-80	70	F8ADC1	356	Peptidase M50 (IPR008915) (AAs: 7 - 127 & 191 - 339) PDZ domain (IPR001478) (AAs: 112 - 193)	6 – 24 (19 aa), 31 – 48 (18 aa), 94 – 116 (23 aa), 227 – 247 (21 aa), 277 – 299 (23 aa), 327 – 345 (19 aa)
Moderate Thermophiles (45-64°C)	Archaea	Picrophilus torridus	Picrophilaceae		60	Q6L2H4	357	None predicted.	86 – 108 (23 aa), 120 – 140 (21 aa), 160 – 178 (19 aa), 190 – 212 (23 aa), 246 – 266 (21 aa), 278 – 298 (21 aa), 304 – 324 (21 aa), 304 – 324 (21 aa), 336 – 355 (20 aa)
Moderate Thermophiles (45-64°C)	Bacteria	Candidatus Desulfofervidus	Candidatus Desulfofervidaceae	50-70	60	A0A127AP35	357	Peptidase M50 (IPR008915) (AAs: 6 - 126 & 190 - 338) PDZ domain (IPR001478)	20 – 37 (18 aa), 57 – 83 (27 aa), 95 – 116 (22 aa), 122 – 145 (24 aa), 157 – 177 (21 aa)

								(AAs: 111 - 189)	
Thermotolerant/ Mesophilic (20-40°C)	Archaea	Candidatus Methanomethylophilus	Methanomassiliicoccaceae	25-40	37	A0A0W7TKD2	183	None predicted.	20 – 37 (18 aa), 57 – 83 (27 aa), 95 – 116 (22 aa), 122 – 145 (24 aa), 157 – 177 (21 aa)
Thermotolerant/ Mesophilic (20-40°C)	Bacteria	Bellilinea caldifistulae	Anaerolineaceae	30-40	37	A0A0N8GNI0	364	Peptidase M50 (IPR008915) (AAs: 48 - 118) CBS domain (IPR000644) (AAs: 235 - 358)	12 – 35 (24 aa), 41 – 59 (19 aa), 99 – 122 (24 aa), 142 – 159 (18 aa), 190 – 217 (28 aa)

Table 9: Information of predicted three-dimensional structure and its accuracy

Class	Group	AC_No (Uni-Prot/ NCBI)	No. of AA	Swiss Model - Template	Seq Identity %	Query Coverage	Zn Binding AA	Quality Check			
								GMQE	QMEAN	MolProbity	Ramachandran Favoured
Hyperthermophiles (>80°C)	Archaea	A0A127BB69	264	4il3A	22.55	5-253	His116, His120, Glu178	0.56	-4.82	1.68	88.66%
Hyperthermophiles (>80°C)	Bacteria	E8T5U6	426	3wklA	35.84	115 - 289	Not Predicted	0.26	-0.82	2.4	89.60%
Extreme Thermophiles (65-79°C)	Archaea	C6A010	412	3b4rB	25.00	168 - 336	His186, Glu187 His190, Asp324	0.17	-7.93	2.83	83.23%
Extreme Thermophiles (65-79°C)	Bacteria	F8ADC1	356	3wklA	32.32	122 - 220	Not Predicted	0.15	-2.97	1.79	89.69%
Moderate Thermophiles (45-64°C)	Archaea	Q6L2H4	357	3b4rB	22.40	87 - 321	His141, His145, Asp271	0.34	-8.72	2.26	84.76%
Moderate Thermophiles (45-64°C)	Bacteria	A0A127AP35	357	2zpmA	32.14	125 - 212	Not Predicted	0.11	-1.52	2.11	89.53%
Thermotolerant / Mesophilic (20-40°C)	Archaea	A0A0W7TKD2	183	3b4rB	31.86	25 - 154	His38, Glu39, His42, Asp148	0.39	-8.83	2.41	81.64%
Thermotolerant / Mesophilic (20-40°C)	Bacteria	A0A0N8GNI0	364	3b4rA	40.36	3 - 233	His59, His63, Asp162	0.46	-4.77	1.8	93.20%

4.CONCLUSION

Thermophiles live in extreme environments of high temperatures. Thermostable proteases of the organisms living in these extreme conditions are of great interest. The comparative studies provide various conclusions. Properties of the thermostable proteases and its comparative studies with other extremophilic proteases provide an estimation of which amino acids are probably responsible for protein stability. Some amino acids of these studies resemble with literature studies while other

amino acids were completely contradictory with the literatures. The results of the comparative studies between thermophilic and mesophilic counterparts shows that alanine and leucine are relatively in greater extent in thermophiles, a trend which contradicts the literature reports. The domain and structural analysis reflected that many predicted structures are membrane bound and depend on the metals for their optimum activity and stability. The activity and stability of the proteins need to be further explored with the help of advanced computational tools to understand the structure and function relationship towards promoting the applications of the proteases.

ACKNOWLEDGEMENT

Authors thanks Gujarat State Biotechnology Mission (GSBTM) to provide computational resources to carry out current studies.

CONFLICT OF INTEREST

No conflict of interest related to the study

REFERENCES

1. Takai K, Nakamura K, Toki T, Tsunogai U, Miyazaki M, Miyazaki J, et al. Cell proliferation at 122 C and isotopically heavy CH4 production by a hyperthermophilic methanogen under high-pressure cultivation. Proceedings of the National Academy of Sciences.2008;105(31):10949-54.
2. Stetter KO. History of discovery of the first hyperthermophiles. Extremophiles.2006;10(5):357-62.
3. Stetter KO. Extremophiles and their adaptation to hot environments. FEBS letters. 1999;452(1-2):22-5.
4. Oda K.New families of carboxyl peptidases: serine-carboxyl peptidases and glutamic peptidases. J Biochem. 2012;151(1):13-25.
5. Endo S.Studies on protease produced by thermophilic bacteria. J Ferment Technol. 1962;40:346-53.
6. Mizusawa K, Yoshida F. Thermophilic Streptomyces Alkaline Proteinase II. THE ROLE OF A SULFHYDRYL GROUP AND THE CONFORMATIONAL STABILITY. Journal of Biological Chemistry. 1973;248(12):4417-23.
7. Zamost BL, Nielsen HK, Starnes RL. Thermostable enzymes for industrial applications. Journal of Industrial Microbiology & Biotechnology. 1991;8(2):71-81.
8. Zhou X-X, Wang Y-B, Pan Y-J, Li W-F. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. Amino acids. 2008;34(1):25-33.
9. Chakravarty S, Varadarajan R. Elucidation of determinants of protein stability through genome sequence analysis. Febs Letters. 2000;470(1):65-9.
10. Lai M, Topp E. Solid-state chemical stability of proteins and peptides. Journal of pharmaceutical sciences. 1999;88(5):489-500.

11. Watanabe K, Hata Y, Kizaki H, Katsube Y, Suzuki Y. The refined crystal structure of Bacillus cereus oligo-1, 6-glucosidase at 2.0 å resolution: structural characterization of proline-substitution sites for protein thermostabilization 1. Journal of molecular biology. 1997;269(1):142-53.
12. Kumar S, Nussinov R. How do thermophilic proteins deal with heat? Cellular and Molecular Life Sciences CMLS. 2001;58(9):1216-33.
13. Neves C, da Costa MS, Santos H. Compatible solutes of the hyperthermophile Palaeococcus ferrophilus: osmoadaptation and thermoadaptation in the order thermococcales. Appl Environ Microbiol. 2005;71(12):8091-8.
14. Das R, Gerstein M. The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics. 2000;1(1):76-88.
15. Shiraki K, Fujiwara S, Imanaka T, Takagi M. Conformational stability of a hyperthermophilic protein in various conditions for denaturation. 2001.
16. Gupta M. Thermostabilization of proteins. Biotechnology and applied biochemistry (USA). 1991.
17. Farias ST, Bonato M. Preferred amino acids and thermostability. Genet Mol Res. 2003;2(4):383-93.
18. Fukuchi S, Nishikawa K. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. Journal of molecular biology. 2001;309(4):835-43.
19. de La Tour CB, Portemer C, Nadal M, Stetter K, Forterre P, Duguet M. Reverse gyrase, a hallmark of the hyperthermophilic archaebacteria. Journal of bacteriology.1990;172(12):6803-8.
20. Roche RS, Voordouw G, Matthews BW. The structural and functional roles of metal ions in thermolysin. CRC critical reviews in biochemistry. 1978;5(1):1-23.
21. Cowan DA, Daniel R. Purification and some properties of an extracellular protease (caldolysin) from an extreme thermophile. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology. 1982;705(3):293-305.
22. Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiology and molecular biology reviews. 2001;65(1):1-43.
23. Taylor TJ, Vaisman II. Discrimination of thermophilic and mesophilic proteins. BMC structural biology. 2010;10(1):S5.
24. Elcock AH. The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. Journal of molecular biology. 1998;284(2):489-502.
25. Zehfus MH, Rose GD. Compact units in proteins. Biochemistry. 1986;25(19):5759-65.

26. DiTursi MK, Kwon S-J, Reeder PJ, Dordick JS. Bioinformatics-driven, rational engineering of protein thermostability. Protein Engineering, Design and Selection. 2006;19(11):517-24.
27. Vakhariya Sakina S JJG, editor Curcumin: A multi-tasking molecule. Proceedings of 9th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952123); 2016: Christ Publications, Rajkot.
28. Ukani H, Purohit MK, Georrge JJ, Paul S, Singh SP. HaloBase. Development of database system for halophilic bacteria and archaea with respect to proteomics, genomics and other molecular traits. J Sci Ind Res. 2011;70:976-81.
29. Rutvi Chovatiya JJG, editor Identification of potential phytochemical inhibitors for the treatment of allergic asthma from the medicinal plants. Proceedings of 9th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952123); 2016: Christ Publications, Rajkot.
30. Rija George ST, Sarah Jacob,, Georrge JJ, editors. Approaches for novel drug target identification. Proceedings of International Science Symposium on Recent Trends in Science and Technology (ISBN: 9788193347553); 2017: Bharti Publications, New Delhi.
31. Nishita NV PS, John J. Georrge, editor Modeling mutations, docking, primer and probe designing of Cytochrome P450 2D6, a drug metabolizing enzyme. Proceedings of 9th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952123); 2016: Christ Publications, Rajkot.
32. Manal Abouelwafa JJG, editor Ebola virus and its potential drug targets. Proceedings of International Science Symposium on Recent Trends in Science and Technology (ISBN: 9788193347553); 2017: Bharti Publications, New Delhi.
33. Kotadiya R, Georrge JJ. In silico approach to identify putative drugs from natural products for Human papillomavirus (HPV) which cause cervical cancer.Life Sciences Leaflets.2015;62:1-13.
34. Joseph V, Georrge J, Pandya J, Jadeja R. O-Vanillin and Some of its Novel Schiff Bases: A Cheminformatic Approach to Identify their Biological Functions. J Theor Comput Sci. 2015;2(136):2.
35. Georrge JJ, Umrania V. Subtractive genomics approach to identify putative drug targets and identification of drug-like molecules for beta subunit of DNA polymerase III in Streptococcus species. Applied biochemistry and biotechnology. 2012;167(5):1377-95.
36. Georrge JJ, Umrania V. In silico identification of putative drug targets in Klebsiella pneumonia MGH78578. 2011.
37. Georrge JJ, editor A Bioinformatics Approach for the Identification of Potential Drug Targets and Identification of Drug-like Molecules for Ribosomal Protein L6 of Staphylococcus species.

Proceedings of 9th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952123); 2016: Christ Publications, Rajkot.

38. Georrge JJ. Docking studies, ADMET prediction of phytochemical inhibitors for Alzheimer's disease. Proceedings of 8th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952116). 2: Christ Publications, Rajkot; 2015. p. 115-20.
39. Gauravi Trivedi JJG, editor Identification of novel drug targets and its Inhibitors from essential genes of human pathogenic Gram positive bacteria. Proceedings of 9th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952123); 2016: Christ Publications, Rajkot.
40. Gauravi Trivedi JJG, editor Bacteriocin producing bacteria from gut of Apis mellifera. Proceedings of 9th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952123); 2016: Christ Publications, Rajkot.
41. Farida Chawala JJG, editor Often Cited Syntactic & Grammatical Errors in Scientific Research Papers. Proceedings of 9th National Level Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952123); 2016: Christ Publications, Rajkot.
42. Consortium U. UniProt: a hub for protein information. Nucleic acids research. 2014;43(D1):D204-D12.
43. Stetter KO, editor History of Discovery of the First Hyperthermophiles. Proceedings of International Symposium on Extremophiles and Their Applications International Symposium on Extremophiles and Their Applications 2005; 2007: Extremobiosphere Research Center, JAMSTEC.
44. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26(5):680-2.
45. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. The proteomics protocols handbook: Springer; 2005. p. 571-607.
46. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 2017;45(D1):D190-D9.
47. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46(W1):W296-W303.
48. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011;27(3):343-50.

49. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 1):12-21.
50. Chen VB, Wedell JR, Wenger RK, Ulrich EL, Markley JL. MolProbity for the masses-of data. J Biomol NMR. 2015;63(1):77-83.
51. Pryor EE, Jr., Horanyi PS, Clark KM, Fedoriw N, Connelly SM, Koszelak-Rosenblum M, et al. Structure of the integral membrane protein CAAX protease Ste24p. Science. 2013;339(6127):1600-4.
52. Quigley A, Dong YY, Pike AC, Dong L, Shrestha L, Berridge G, et al. The structural basis of ZMPSTE24-dependent laminopathies. Science. 2013;339(6127):1604-7.
53. Rawlings ND, Barrett AJ. Evolutionary families of metallopeptidases. Methods Enzymol. 1995;248:183-228.
54. Ponting CP, Phillips C, Davies KE, Blake DJ. PDZ domains: targeting signalling molecules to sub-membranous sites. Bioessays. 1997;19(6):469-79.
55. Ponting CP. Evidence for PDZ domains in bacteria, yeast, and plants. Protein Sci. 1997;6(2):464-8.
56. Scott JW, Hawley SA, Green KA, Anis M, Stewart G, Scullion GA, et al. CBS domains form energy-sensing modules whose binding of adenosine ligands is disrupted by disease mutations. J Clin Invest. 2004;113(2):274-84.
57. Kemp BE. Bateman domains and adenosine derivatives form a binding contract. J Clin Invest. 2004;113(2):182-4.
58. Janosik M, Kery V, Gaustadnes M, Maclean KN, Kraus JP. Regulation of human cystathionine beta-synthase by S-adenosyl-L-methionine: evidence for two catalytically active conformations involving an autoinhibitory domain in the C-terminal region. Biochemistry. 2001;40(35):10625-33.
59. Zhang R, Evans G, Rotella FJ, Westbrook EM, Beno D, Huberman E, et al. Characteristics and crystal structure of bacterial inosine-5'-monophosphate dehydrogenase. Biochemistry. 1999;38(15):4691-700.
60 Bethesda (MD). Entrez Help. 2016 [Available from: https://www.ncbi.nlm.nih.gov/books/NBK3837/].