CHAPTER 3

EVOLUTIONARY ANALYSIS
## CHAPTER 3

### Table of Contents

Chapter 3: Evolutionary Analysis

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Name of the Sub-Title</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>Evolutionary basis of sequence alignment</td>
<td>41</td>
</tr>
<tr>
<td>3.3</td>
<td>Phylogenetic analysis</td>
<td>41</td>
</tr>
<tr>
<td>3.4</td>
<td>Background study</td>
<td>44</td>
</tr>
<tr>
<td>3.5</td>
<td>Comparative study of yeast gpl12 and Mammalian Pig L</td>
<td>46</td>
</tr>
<tr>
<td>3.6</td>
<td>Lipophosphoglycans and phosphatidyl inositol Mannosides</td>
<td>47</td>
</tr>
<tr>
<td>3.7</td>
<td>Materials and methods</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>3.7.1 Selection of sample protein structure</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>3.7.2 Multiple sequence analysis</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>3.7.2.1 Clustal</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>3.7.3 Phylogenetic study</td>
<td>57</td>
</tr>
<tr>
<td>3.8</td>
<td>Results and discussion</td>
<td>59</td>
</tr>
<tr>
<td>3.9</td>
<td>Discussion</td>
<td>81</td>
</tr>
</tbody>
</table>
3.1 INTRODUCTION

In the early days of sequence analysis of proteins and genes[103][104], The sequences from related protein sequences and gene sequences were similar. Bioinformatics based on The discovery was imperative as it shows the high similarity between two genes is a well-built argument for their homology. The basis for comparison is mainly related to the evolution as they have a common ancestor. Random mutations In the sequences random mutation takes place build up in due course of time so that proteins that have an extensive ancestor as proteins or genes that diverged from each other more recently. In sequence alignment give interesting information that which parts of the sequence is significant for the function. In a protein sequence, sequence alignment has become an essential part of biological science [105].

3.2 EVOLUTIONARY BASIS OF SEQUENCE ALIGNMENT

Sequence alignment is when two sequences show similarity it is called homology of sequences [106] [107]. Although, these two terms often interchanged in modern usage, let us distinguish them to avoid confusion in the current discussion. The similarity is a percent identity. Homology is two genes share a common ancestor [108].

3.3 PHYLOGENTIC ANALYSIS

This analysis infers the evolutionary relationships among genes from across or within distinct populations of different species. It takes
aligned genetic sequences as input; it is to work out the pattern of ancestry. This analysis plays a significant role in drug or vaccine development [109] [110]. Used methods of phylogenetic analysis fall into four categories such as distance-based methods, maximum parsimony, maximum likelihood and Bayesian methods [111][112].

Distance-based methods such as data clustering methods that reflect on only the pairwise measure of evolutionary distances between sequences. Whereas maximum parsimony presumes the correct phylogenetic tree is the one involve the smallest number of evolutionary measures to explain the input sequences. This method requires a substitution model to assess the probability of particular phylogenetic trees that aims to discover a tree with the highest likelihood with respect to the given substitution model. However, Bayesian methods rely on probabilistic models of sequence evolution similar to maximum likelihood tree inference. Unlike from maximum likelihood estimates, still, trees are constructed for case based on the consensus of a set of trees sampled from the highest probability regions of the posterior distribution over evolutionary model parameters as well as trees. These methods are considered to be the most accurate and relatively high-speed due to the computational capacity of modern computers and methodological advances [113][114][114]. On the contrary, these methods depend on large numbers of parameters that have to be estimated. These include branch lengths, tree topology, parameters of the substitution model and site-specific rate variations that make them even more
computationally demanding. UPGMA and NJ algorithm are quite fast with runtime depending only quadratically and cubically, respectively, on the number of input sequences [116][117]. Conversely, these are less accurate than all other techniques. Here in this study, the focus is on maximum parsimony standard for tree inference. While, this approach has been a reasonable trade-off between speed and accuracy and is still considered as an important optimality criterion for the evaluation of phylogenetic trees [118], especially for data sets at lower evolutionary divergence [113].

GlcNAc-PI-de-N-acetylase enzyme commonly referred to as GPI12P in prokaryotes and eukaryotes is highly co-related to PIG-L enzyme of mammalian cell’s GPI system based on its functional analysis [119]. In fact, the two terms GPI12P and PIG-L are most often used synonymously to refer to the enzyme that involved in the deacetylation process of GPI anchor Biosynthesis [120]. The sequence of reactions and most of the endogenous donors of the anchor components is now fairly well defined in most systems. As the focus of the investigation shifted to specific reactions in the pathway, it has become apparent that there are significant differences in GPI biosynthesis in different species. These differences are well defined between trypanosomes and mammalian cells and have been suggested between yeast and mammalian cells. Therefore, specific steps in the GPI biosynthetic pathway are potential targets for anti-parasitic and anti-fungal agents. Future investigation of the enzymes and reaction
mechanisms of GPI biosynthesis will facilitate the development of inhibitors that could be used as drugs.

These inhibitors may also be useful for elucidating the functional area of the GPI anchor for various proteins in different species or cell types [121]. This active site chemistry will help us identify the diversity among the GPI deacetylases and Histone deacetylases superfamily enzymes. Hence, the idea is to construct a phylogeny of all the proteins. Any phylogeny primarily shows the relatedness between the given sequences in terms of amino acids linked by a chain via the peptide bonds.

3.4 BACKGROUND STUDY:

On an overview, the biosynthesis of GPI is initiated on the cytoplasmic side of the ER by the transfer of N- acetyl glucosamine to phosphatidylinositol (step1). GlcN-PI and/or GlcN-acyl-PI precursors (steps 2-3) translocate from the cytoplasmic side to the luminal side of the ER by a putative flippase (step 4). Steps 2 to 6 occur in the mitochondria-associated membranes. GPI is transferred to carboxyl terminal (step 11). GPI-anchored proteins are transported to the cell surface [122][123].
Fig. 3.1: Process of attachment of GPI anchor to the Protein by deacetylating C-terminal peptide

Fig. 3.2: Comparison of several GPI structures

In the above figures 3.1 and 3.2, the structures are shown to highlight the conserved core regions and to indicate differences found in various anchors, acetyl cholin esterase. Components shown are ethanol amine (pink rectangles), phosphate (black circles), mannose (green circles), glucosamine (blue circles), inositol (purple hexagons), galactose (Light blue herons) and N-acetyl galactosamine (orange hexagons). Fatty acyl and alkyl chains are as follows: red, myristate; 1, acyl (22:4, 22:5, 22:6); 2, alkyl (I&O, 18~1); 3, acyl (16:0). Where not identified they are not known. Deacetylation
is known to be a prerequisite for the subsequent mannosylation reaction. The de-N-acetylation step is catalyzed by the product of PIG-L/ Gpi12 gene in both eukaryote and bacteria [33]. Several PIG-L proteins have been identified as slightly similar in eukaryotes, Archea, and Bacteria. The yeast PIG-L protein, also known as GPI12, is known to be essential for cell viability, as is the mammalian homologue. GPI 12 shows 24 % identity with the rat PIG-L [33]. Human Pig-L protein is a 252 amino acid showing 77% homology with rat PIG-L [124]. Unlike GPI12, this protein has been shown to possess two ER localization signals and has been found to be localized on the ER and in the mitochondrial associated ER- membrane. Not merely confirmation, there appear to be species-specific differences in PIG-L proteins. For example, PIG-L of Trypanosoma enzyme, unlike its human counterpart recognizes and catalyzes the Deacetylation of GlcNAc [33].

Also, GPI proteins, GPI anchors are found in several species. However, signaling peptides are released through the activity of enzymes [125]. In some species, this function is absent due to lack of signaling gene.

3.5 COMPARISON STUDY OF YEAST GPI12P AND MAMMALIAN PIG-L

The second step of GPI biosynthesis is de-N-acetylation of GlcNAc-PI. According to Watanabe et al. [120], the recombinant rat PIG-L protein purified from Escherichia coli as an involved with GroEL
has GlcNAc-PI de-N-acetylase activity in vitro. The Saccharomyces cerevisiae YMR281W open reading frame encodes for GPI12p protein that is only 24% similar to PIG-L protein in mammals. It was reported that in mammalian cells that are deficient with PIG-L, and when they are transfected with the Yeast GPI12p, the cell restored its de-N-acetylation activity. So it can be ascertained that Yeast GPI12p and Mammalian PIG-L are only 24% similar in sequence but functionally conserved throughout the evolution [123]. In other words, though their chemical structure differs considerably, they retain the same functionality.

Phylogenetic Studies on GPI-12p and its counterparts i.e. isoenzymes identified from then other organisms is performed to understand the functional character relatedness amongst organisms. Since, genetic relationships between species can be represented using phylogenetic trees; that, mainly reveals the differences in their sequence or other words the chemistry of the protein isozymes. Several computational methods are available to build the trees. With this method, these trees can be constructed by comparing sequences of various species. A relationship between the species is classified as a phylogeny.

3.6 LIPOPHOSPHOGLYCANS AND PHOSPHATIDYL INOSITOL MANNOSIDES

Cell surface pathogens are involved in host-pathogen interactions [126]. These are based on a type-2 GPI core, Man α1-
3Man α1-4GlcNa1-6PI, repeat domain carry modifications [78]. In Saccharomyces cerevisiae, the component of lipid is a monoalkyl-lysophosphatidic inositol with alkyl groups at C22 and C24.

Further related lipids are the phosphatidylinositol mannosidase [127]. The lipids range in from simple mono-mannosidases in Saccharomyces cerevisiae.

![Phosphatidylinositol Dimannoside Structure](image)

Fig. 3.3: phosphatidylinositol dimannoside structure

The phosphatidylinositol di mannoside from Saccharomyces cerevisiae [127]. The principal fatty acid constituents are palmitic and 10-methyl-stearic (tuberculostearic) acids.

The lipomannan's have a longer chain of mannose units, and these can be further modified with arabinan to produce the highly complex lipoarabinomannan [128]. A metal ion induced enhancement in enzymatic activity.
Knowledge of the biosynthetic enzymes leads to improving drug therapies. Further, structural similarities can also point us directly to the functional analysis, especially if the active site structures are matched. Therefore in this study, the relationships between the structures are estimated according to the sequence matches of amino acids in the proteins using the phylogenetic software. The understanding chemical diversity of deacetylasies that are involved in membrane anchoring across species and also across evolution using Phylogenetic analysis. Homology modeling of GPI12P from known similar deacetylase structures as templates to study the active site and thereby understand the function of protein

3.7 MATERIALS AND METHODS

Sequences were retrieved from NCBI [129] and Swiss-Prot datasets [130]. Conducted BLASTP [131] search against the DAD database [132] (all amino acid sequences) on DNA Databank of Japan (DDBJ) [133] to collect full-length sequence of GPI12p and PIG-L protein sequences in eukaryotes and prokaryotes and chose sequences with E-value below 10. From the hits obtained by homology search, taken the sequences that contained GPI12p and PIG-L protein sequences. These sequences were used for further analysis, and NJ phylogenetic tree was constructed. Only the complete sequences excluding the partial sequences were searched in all three comprehensive datasets, Genbank, dbEST and dbGSS of all species.
3.7.1 Selection of Sample Protein Structures:

Sequence retrieval was done and the number of sequences retrieved from BLAST searches and chosen for further analysis. The below table 3.1 gives the names of the GPI12p and PIG-L protein molecules that do the similar enzymatic function, even though, some of them are structurally diverse. There are thousands of sequences in the NCBI database from hundreds of species. Protein structures from NCBI database from various species that are functionally similar to GPI12p and involved in GPI anchor biosynthesis

> EJS44137.1 | Gpi12p Saccharomyces arboricola

MSRLAKVSFKLLYYKITYLAIYIJTYPKITTTRNNSLKHVAHKNSDSQINL VIAHDPDEVMFFSPVISQHLYFPNTPFNIICMSKDAEELGETRVEKLNDSASLL LQNGRPVSVQMDFEDGMDKVEINSITSTLSKTIDLNNEKLNQIIITDFDSYGVSDH INHKSCTAVKLINDYTQSKTEKNEESPHTALYLKSYKNIFLKYSSFIWEILRLYLNVLPSFHKDQVPLPPTITTEQSRLLLTTNTAQYILALAAKLNAHKSQMVWFRYGW WILSRFVYVNDIEVYT

> NP_014008.1 | Gpi12p Saccharomyces cerevisiae

MKMLRRTKVNFSKLLYYKITYLAIYIJTYPKIVSRNNASLQHIFFHPKYGDEI NLVIAHDPDEVMFFSPIISQLNSYPRTVFNPNIICLNGAEGLGETRVELNESAA LLLHNERAVSVQVMDQGDMEIWDIDSITSSLSQKIDIKNHNQNIVTFDSYGVSNHINHKSCYAAVKKLVDDYAQPPTKRNEQPPHVTALYLR SYKNNIVALYNSFIWEILKILYDLISPFRRIIQALPPNTAEEKDLKLSMNTAQYVLAFATMLNAHESQVVFWRGWWIFSRFVYVNDIEVYT

> EXX61259.1 | Gpi12p Rhiophaghusirregularis
MISFVIYSLFIFLMFLITNTIILYYIMINNNEVQLNERKILLIISHPDDECMFFGPT
LLFLRNRRKQNQVHLLCLSIGNESGLGEIRRNELESSCITLGDIEINIIDHPSLQDG
PKNNWDPLTISNIINDYVTKHKIDIITTFDEKVGSHPHNAALNGARHFIQSVTTS
NDIILYKLLTPVLRYISILDLPLTYLFKNRNNNLHIFISSFTNFRARKAMTSH
QSQLVWFHRHYILFSYMMINELQLVNIN

>AFA76121.1 Actinomycetales—Gordoniapolyisoprenivorans

MTVPVRMLVHAPDDSLWTGTTIARHVELGGDVSVVTATWATGTSRHGELRNALE
LGVTREPIMLGFADDGVPSAPGAKRFCVDFSDKQVRILVGHIRRLRPEILLTYDPV
GIYGHRDHVHAHRRLAASADAAGASMLHRSTGFAWRISVYMATIADWMIEDVDDF
FPTISRYALPGTSPGIDLELTDVSAWVKQKAACAVASHREVERSQMISSLMNPRER
RDRLFGTECYIRRLDVRGVDL

>ERL61453.1g—proteobacteria—Piscirickettsiasalmonis

MIKFSEPGADLFIPIGGGSIANLSRSTSHLAIAAHQQDIEILAFHIGAECYQAERA"GF
CGVVVTDGASPRAGDYAETYTDKKMQEQVRAEETQRAAKLGEYSAQFQLAYPSH"QVKN
TSALPLVEDELEELIKTQPEVYVLHNPMDKHDTHVAVFLKSLAALRRLPEKIRPKV"V
YACEVWRLDLWVCDKIELKIAKDEPRELALALLQVFDSQVSGGKRYDQAVLGR"LAH
ATFSSHSHAVDDVKVGYAIDVMPLLENPMLTVEDYVRGFVERFAHSLTTTLT
ALSQD

>AIK33361.1 Bacillales—B.anthracis

MKNERHVILVFPDPDESYCVAVTILAYTQRNVPLTYVCLTLGEMGRAMGNNPPFAT"R
ESLYAIREKELKATNILGKDLRMGMYRDKTEFEFGELRVIQKCVELNPSLV
ISFYPGYAVHDHATGEAVAEALATIPNEKRPTFYAVAFANNHEAEI"GPPP"HV"KNEV
KEYVPKKLEALQAHASQFATKVTEKREYEDGTETEVLEPREPFIYPFKDNK

>EKY07763.1 CFB group—Capnocytophaga sp.

MKKVIVSAHPDDEILGAGTLLHKHKKNGDIYWLIVTNVlescGSKERCVSRQKE
IEKISEALGVEKVFLLNYPTMSLSTSTLIEMPKISKVFTEIEPEIVYCLNRSDAHS
>WP_020466630.1 Plantomycetes-Singulisphaeraacidiphila
MLSAAKGRGTHGHLSDVRESLRQAVGRDRLPVMLCPHADDGAISAACLLHEYAVRRG
LPVIEVLVFAGERNVAAAPWLNDQKKVTRVRESEFRLECSVLGAEAVCWDLAYRSPGY
QPSAGDIDKIVEWFIQRRPAGAILPPATDAHVAHRMTRALASIGLVGQQLKDTLVLT
GWTPGWGPLPQPYNAYSFYSDGAEARTKEWAICHASVQLTDYTKYCSHLGRAYAALTR
EWAEGHSLTGRAARTDDRFIGVELFQIEKYDPELDASATDPLQIALGILSGH
VAPASLHAIATTV

>WP_031460266.1 GNS bacteria-Chloroflexus sp.
MRPTLLAVFAHPDDESFGPGTTLARYAWSGAVAHICATGGEEGTVDAELLRGYDSV
AALRRTELLRAAEALGLSSTTLLGYRDSGMPGTEANRHPAALINQPRERVVQQLVRL
MRQLRPQVVITFDPGIGYRYPDHIAIHEATVAAFAAAGDPTFPEAGSAYAPQKLYY
TTFSSRRLLLVLKLMPLFGRDPRAFGRNRDIDLMQLVAVDFFTHARIDTTAVQAQAH
AAALAHVSQGAGNPLIRSLLGKVANLLGRSDTFMRAYPPATSNVREDDLFTG

>WP_027840276.1 Cyanobacteria-Mastigocoleustestarum
MSKLTLKNRLQIFTGEVVNRINSKVVLFTAKFKSRIKANNKTAIVFSHPQDESLLG
CGGMIAIKRSLGVSVKVVFTGDRYGRPDWIEPEDIIQFQKEATLNILGVEQSD
TYFLGEDPSQLSESQRYDLISRKLKSSRPEEVYYPHQDGDHPDHEATFDLV
LAAIAHSGLETLEFQYPIWVFQWNPQLNKLHRENLTNAHRVIAAVKKQKQAIESY
KSQLRYLPYGVALLSNSEEIFFK

>EEB06371.1 Fungi Schizosaccharomycesjaponicus
MNWLLITIAVSIAGITFALNAASPGHQLLDHDLVVFVAHPDDESMMGFPTVDFVSR
RRHGGVHLLCLSNGANDGLGAIKEIVEAASYKIPASNVHVVEDPKLQDGMRNTW
STTAVASAVGSAIESAKSAGRISISLVIITFDGYGISGHANHRCHRGRVIDYAREHRL
RVYTLDSVNLKRYWTLLLDATFTYLTLYRWGSERRVVCADHRAQARI
3DAMVEAHK
QMVWYRQIWISLSRYMSTNSLRR

> KFX49055.1 Fungi - Talaromyces marneffei

MPMYKPLLTAATHLTRRAAKRMSIISILLISLFPITLYLLGNLNLVDPLR
AIRNAQNLFFITAHPDGALFFGPIKHPETHKLETKSFCLEYSIPKDRCLILQNLKD
QFSKKWDESIIQIRLERHIRKWEIDLIITFDADGFDGDNHRSLAVQRYSTVHDH
NPTAFGDDFatQWDQDFFGVCSCAGGGVSDPLDGKRVGVQPPNGDAYGD
DWSGHGKQAALKHKSAYSWDRAVQSSLSRYMWFDLRLRM

> BAA74775.1 Animalia - Homo sapiens

MEAMWLCVAVAVLAVGFLVWDSRSMKSEQQGLGAEFLTLLVIAHPDDEAMFF
APTIVGLARLRHVVVLCCSFAGNYNQGETRKELLQSCVDLGLPSSVMIDNDF
PDDPGMQWDTEHVARVLQHIENVINLVFTDAEVGSGHSHIALLYAAVRALHSEG
KLPKGCSVLTLQSVNLKRYISLLDLPLSSLHTQDVLFLVESKEVAQAKKAMSCHRS
QLWFRFRLYIFRSYMRINSLSFL

> BAA20869.1 Animalia - Rattus norvegicus

MEVVGGLCCVAVAVLAVGLRVNSAERMRSPEQAGLPAGSRALVIAHPDDEAMFF
APTILGLARLQQSTCCSFSGNYNQGEIRKELLQSCAVLGIPPSVMIDKREF
PDDPEQWDTEHIVSTILHIAHIATDLVTFDAEVGSGHSHIALLYKAVRALHSEG
KLPKGCSVLTLQSVNLKRYVFLDLPLSSLHPQGFLVFLESKEVAQAKKAMSCHRS
QLWFRHLYTFVRMSVNSLQLL

> AEE79746.1 Plantae - Arabidopsis thaliana

MAWLVVSVALIVWLSASCFKIFRATISGAIILDDGKTPQKNNLVFLVIAHPDDES
FFSPINTYNLASACNHLMCLSTGNADGMSIRNNHELACAVLVKPLQQLKILNH
NLQDGFAQWSHDLEISTEEEETKHHICTITTDNYGVGQHCIRHHDVHRGFLQ
TNSGRNVAWELVSLNIFRKYCGPVIDWLSILSAHIHPKVIIIENEQPWKSFKAM
HLSQWVWWFRKLFVFSFSSYANTSLRIN
>AFW59914.1 Plantae—Zea mays
MAWIWMLAGAVLLWAISLGRVLSAAPSCVPSSPQFMPLSGDRRSRNVLVVAH
PDDESMFFAPITLFLKSKGSHIHLCRGNADGLGDTRKEEELYHACVSLKIPHEQV
KVLDPKQLGDFHKEKWDHGIDLVAELMHEVQLWAIDTIVFDSYGVGHPNHKDVHHG
ICKFLHTNRQGNVEAWELASLIRKYGSPIDIWLSLTSFAWTKQIYTVLNNSSPS
RSYEAMAAHRSQWVFRRLFVMSSITYINMLRR

>AHB67707.1 Archea—Haloarculahispanica
MDVLVVVAHPDDADVFCGGTIHKAERGDEVSIHVMTIGYEYGGLRTDSQEAVGRVRE
QEARASGAVLGSVEVALEFKDGRITYSLENRMVEMDVIREYDPDILTHYKDDLHP
DHRATSRLVTDAYMASLPLVETDFEPDPDNYYFGKPTSEFTPSMFIDIDGYLEQ
KVTAIKKHEVSQVEFLVEHGGIDAEFDNLIDGLRAENLFGKQAGRSAEGFVPLHES
AQEFLG

Table 3.1: Protein structures from NCBI database from various species that are functionally similar to GPI12p and involved in GPI anchor biosynthesis

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIGL</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>CG4433</td>
<td>Drosophila melanogaster</td>
</tr>
<tr>
<td>Pigl</td>
<td>Rattus norvegicus</td>
</tr>
<tr>
<td>Pigl</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>PIGL</td>
<td>Gallus gallus</td>
</tr>
<tr>
<td>PIGL</td>
<td>Pan troglodytes</td>
</tr>
<tr>
<td>PIGL</td>
<td>Canis lupus familiaris</td>
</tr>
<tr>
<td>PIGL</td>
<td>Bos Taurus</td>
</tr>
<tr>
<td>Pigl</td>
<td>Danio rerio</td>
</tr>
<tr>
<td>AT2G27340</td>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td>AT3G58130</td>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td>GPI12</td>
<td>Saccharomyces cerevisiae</td>
</tr>
<tr>
<td>AgaP_AGAP001955</td>
<td>Anopheles gambiae</td>
</tr>
</tbody>
</table>
3.7.2 Multiple Sequence Alignment:

To achieve the maximum, several approaches of aligning of sequences have been performed. A static assignment of homologous
sites is produced by progressive multiple alignments using Clustal tools [134] in three steps. The process of creating multiple alignments begins with computing all pairwise alignments, which is followed by constructing a dendrogram. It defines the groupings of the sequences by similarity, and it is used as a guide tree for the order of sequences aligning to carry out the final multiple alignments.

3.7.2.1. Clustal

Clustal is a widely used multiple sequence alignment computer programs [134]. One is Clustal W: This is Command line interface, and the other one is Clustal X: This program is available from the Clustal Homepage. ClustalW2 is a multiple sequence alignment program for DNA or proteins. These steps of the program are in the below Figure 3.5.

Fig. 3.5: Clustal Home Page.
3.7.3 Phylogenetic Studies

Phylogenetic methods used to study the evolutionary relationships of organisms using phylogenetic trees to compare species and used to analyze [111]. The important application is to test for phylogenetic tendency [135].

Table 3.2: Protein sequences of organisms used in the study

<table>
<thead>
<tr>
<th>PIG-L proteins took for the analysis from various kingdoms spanning the entire Tree of life</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bacteria</strong></td>
</tr>
<tr>
<td>Actinobacteria</td>
</tr>
<tr>
<td>Gordonia polyisoprenivorans</td>
</tr>
<tr>
<td>Proteobacteria</td>
</tr>
<tr>
<td>Piscirickettsia salmonis</td>
</tr>
<tr>
<td>Firmicutes</td>
</tr>
<tr>
<td>Bacillus anthracis</td>
</tr>
<tr>
<td>CFB group bacteria</td>
</tr>
<tr>
<td>Capnocytophaga sp.</td>
</tr>
<tr>
<td>Planctomycetes</td>
</tr>
<tr>
<td>Singulisphaera acidiphila</td>
</tr>
<tr>
<td>GNS bacteria</td>
</tr>
<tr>
<td>Chloroflexus sp.</td>
</tr>
<tr>
<td>Cyanobacteria</td>
</tr>
<tr>
<td>Mastigocoleus testarum</td>
</tr>
</tbody>
</table>

The names, symbols used in databases, GenBank IDs, and species they belong were tabulated below in the table. Their protein sequences were collected in fasta format and stored in the text files for
further bioinformatics analysis towards identifying the similarities amongst the protein structures by first identifying the protein sequences difference.

Table 3.3: Protein structures from NCBI database from various species that are functionally similar to GPI12p and involved in GPI anchor biosynthesis

<table>
<thead>
<tr>
<th>ID</th>
<th>Symbol</th>
<th>Protein</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneID:9487</td>
<td>PIGL</td>
<td>NP_004269.1</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>GeneID:42388</td>
<td>CG4433</td>
<td>NP_650857.2</td>
<td>Drosophila melanogaster</td>
</tr>
<tr>
<td>GeneID:192263</td>
<td>Pigl</td>
<td>NP_620256.1</td>
<td>Rattus norvegicus</td>
</tr>
<tr>
<td>GeneID:327942</td>
<td>Pigl</td>
<td>NP_001034625.1</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>GeneID:417600</td>
<td>PIGL</td>
<td>XP_415845.2</td>
<td>Gallus gallus</td>
</tr>
<tr>
<td>GeneID:468371</td>
<td>PIGL</td>
<td>XP_523760.2</td>
<td>Pan troglodytes</td>
</tr>
<tr>
<td>GeneID:479514</td>
<td>PIGL</td>
<td>XP_536652.1</td>
<td>Canis lupus familiaris</td>
</tr>
<tr>
<td>GeneID:507080</td>
<td>PIGL</td>
<td>XP_583634.2</td>
<td>Bos Taurus</td>
</tr>
<tr>
<td>GeneID:561471</td>
<td>Pigl</td>
<td>XP_689961.2</td>
<td>Danio rerio</td>
</tr>
<tr>
<td>GeneID:817277</td>
<td>AT2G27340</td>
<td>NP_973543.1</td>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td>GeneID:824982</td>
<td>AT3G58130</td>
<td>NP_001030882.1</td>
<td>Arabidopsis thaliana</td>
</tr>
<tr>
<td>GeneID:855324</td>
<td>GPI12</td>
<td>NP_014008.1</td>
<td>Saccharomyces cerevisiae</td>
</tr>
<tr>
<td>GeneID:1281168</td>
<td>AGAP001955</td>
<td>XP_321107.2</td>
<td>Anopheles gambiae</td>
</tr>
<tr>
<td>GeneID:2543490</td>
<td>gpi12</td>
<td>NP_593887.1</td>
<td>Schizosaccharomyces pombe</td>
</tr>
<tr>
<td>GeneID:2675729</td>
<td>MGG_05020</td>
<td>XP_359757.1</td>
<td>Magnaporthe grisea</td>
</tr>
<tr>
<td>GeneID:2711752</td>
<td>NCU09510.1</td>
<td>XP_329869.1</td>
<td>Neurospora crassa</td>
</tr>
<tr>
<td>GeneID:2893020</td>
<td>KLA0D01430g</td>
<td>XP_453132.1</td>
<td>Kluyveromyces lactis</td>
</tr>
<tr>
<td>GeneID:4337416</td>
<td>Os04g0678800</td>
<td>NP_001054275.1</td>
<td>Oryza sativa</td>
</tr>
<tr>
<td>GeneID:4618907</td>
<td>AGOS_ABL120W</td>
<td>NP_982827.1</td>
<td>Eremothecium gossypii</td>
</tr>
</tbody>
</table>

Sequence analysis was done using multiple alignments using Clustal X tool, PSIPRED [136] and MEGA software [137]. Score
matrices were calculated using BLOSUM 62. This score was given as an input to maximum parsimony algorithm for reconstruction of the phylogenetic tree. The MSA alignment of all the GPI12p and PIG-L protein sequences were color coded with CHROMAS software [138] and presented as following.

### 3.8 RESULTS AND DISCUSSION

From about 420 GPI12p and PIG-L protein sequences, only 17 sequences are based on covering the span of the life forms according to the following tree of life construction.

![Tree of Life](image)

Fig. 3.6: Tree of Life constructed radially using complete genomes

A Tree Of Life, constructed, an online phylogenetic tree viewer [139]. Eukaryotes have been colored such as red, archaea green and bacteria blue. Therefore, since these 17 protein sequences of GPI12p and PIG-L are taken from each of the taxa nodes of tree of life, with the advent of simple sample it gives a unique opportunity to identify the similarities of the human PIG-L protein structure with the GPI12p
protein of Saccharomyces cerevisiae along with all other proteins from world life forms. Refer to the below presented table to know the proteins used obtained from each species as represented in each kingdom of life.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJS44137.1</td>
<td>-MSRLAKVFSKLYKIKTLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>NP_014008.1</td>
<td>(4)RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>BAA74775.1</td>
<td>-----MEAMWLLCVAALAVWFLWVMSERMKRESQGQRLGE-------SRTLVI</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>BAA20869.1</td>
<td>-----MEVGLCVAALAVWFLWVMSERMKRESQGQRLGE-------SRTLVI</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>EXX61267.1</td>
<td>-----KJIFVYSLFIPMLFLNINTNTIILYMIINNEVQNL-------RKLILL</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>EEB06371.1</td>
<td>-----MNWLLITIAVIAGTIFALN-------AASPGQHQLL-------HVDFVF</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>AEE79476.1</td>
<td>-----MAWLVVSVLIVWLSFCIFFRATS-----ISGAAILDDGKTPQK-------KVFLVI</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>AFW59914.1</td>
<td>-----AHHWVLAAGAVLWAILSGLRVLSSAASCVPQPMPFSLGDRRS---RNYLVVV</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>KFX49055.1</td>
<td>(16)RAAKRRMSLCLTVSKCIRNRI.........SACVM</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>AIK33673.1</td>
<td>-------------------------------MKNE-------RNYLVVV</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>AHB67707.1</td>
<td>-------------------------------MKVEIVIS</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>WP_027840276.1</td>
<td>-------------------------------MSKTLKRNQIIFTGVRVINNSKVLFLTKFQSRIKAN-------NKTVF</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>WP_031460266.1</td>
<td>---------------------------------------------------------------MRPSTLAVF</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>AFA76121.1</td>
<td>---------------------------------------------------------------MTVPKMLVH</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>BP_020466630.1</td>
<td>---------------------------------------------------------------MLSAAKGRTGHHLSDVRESLRQAVGR-------DRPEVLMLC</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>ERL61453.1</td>
<td>-------------------------------MDDNNNEVQLNE-----------AGDHY</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
<tr>
<td>Consensus/80%</td>
<td>-------------------------------MDDNNNEVQLNE-----------AGDHY</td>
<td>RRTKVNFSKLLYKITLAIVLTLIYIYPKTITTRNNESLHKVFASHKNSDSQNLVI</td>
</tr>
</tbody>
</table>

Consensus/80%
Consensus/80%  

Consensus/80%
AHB67707.1  KGAEGAVFPEHSAQEFLG--
EKY07763.1  ASVVEYAAFIKYIDK------
WP_027840276.1  K------------------
WP_031460266.1  DTRAYPPATSVDSDLFTGVQW
AFA76121.1  RERRDRLEOTECYRRDLVRGGLD
WP_020466630.1  LQALGLSGLVAPAHTATTV--
ERL61453.1  ERFAHSLTLTALSDQ--------
Consensus/80%  b.h....h..b..hpp.........

Fig. 3.7: Results of multiple sequence alignment.

The Multiple sequence alignments were analyzed to understand the conserved sequences, conserved MOTIFs, and check the conserved areas whether they fall into the active site region. Further, the below presented colour-scheme of the alignment is a result of secondary structure type identification.
Fig. 3.8: Image of PSIPRED results
A color represents each sequence in secondary structure. The 3-state (Helices in red & Strands in Blue). If sequence had no color assigned, then there is no information on DSSP available or that no prediction was possible for that sequence. The MSA was analyzed for Secondary structure prediction using DSSP [140] and PSIPRED [136]. Furthermore, the PSIPRED online MSA analyzer tool was used to predict the hydrophobicity areas according to the conserved structures. The hydrophobicity scale used is from Eisenberg et al. [141], which is shown on the right side of the MSA color coded picture according to the level of understanding.

The color assignments from hydrophobic to hydrophilic sequences are as shown in the following MSA picture. The conservation scoring is performed by PRALINE [142]. The scoring scheme works from 0 for the least conserved alignment position, up to 10 for the most conserved alignment position. The Hydrophobic and hydrophilic sequences are crucial in identifying the structural similarities in the protein because a similar trend in the Hydrophobic and hydrophilic scoring will often give similar structures or similar active structures. This profile points to the similarity and dissimilarity of the proteins and thereby their functions can be inferred. In this profile, it manifests to us that throughout the kingdoms of life i.e. animal kingdom, plant kingdom, and others, some conserved sequences along the protein retained the function.
Fig. 3.9: Colour assignments from hydrophobic to hydrophilic sequences.

With this information in hand, Maximum Parsimony method was used inferred a sequence similarity by phylogeny method [143]. The most parsimonious tree with length 1274 is shown. Using the
Close-Neighbour-Interchange algorithm with search level 2 the MP tree was obtained [144] in which the initial trees have been achieved. Phylogenetic analyzes were conducted in MEGA4 [130].

Fig.3.10: Phylogram of 17 selected GPI12p and PIG-L proteins sample from all the kingdoms of lifeforms

Based on the comprehensive phylogenetic analysis conducted with 17 protein sequences of GPI12p and PIG-L type, it is clearly revealed that the human PIG-L protein is shared a high similarity with the Saccharomyces Protein i.e. GPI12p. However, these two proteins of Human and Saccharomyces spp are seemingly highly diverged from proteins of Archea proteins. Even though, Human PIG-L is closer to Arabidopsis spp of Kingdom Plantae compared to any fungi other than Saccharomyces spp. Further to analyze the structural similarities of
the PIG-L and GPI12p proteins, these were compared with the entirely known protein populace histone deacetylase superfamily. These proteins or enzymes are classified into histone deacetylase superfamily based their function. Here in this study the proteins are sorted into three groups as Archea, Bacteria, and Eukaryota. The abbreviations designated for the enzyme types used in the figure 3.10 above and tables below are as follows: Polysaccharide deacetylase (p-d), Histone deacetylase (h-d), Acetoin utilization protein (AUP), Protein Deacetylase (Pr-d), Putative deacetylase (Put-d), SIR2 family (NAD-dependent) (SIR2), Succinyl-diaminopimelate desuccinylase (Sdd).

These classes of enzymes mentioned above are ubiquitous and share almost standard function of deacetylation with the PIG-L and GPI12p proteins. They also share certain LmbE and Putative sequences in common among their sequences. Hence, these sequences are selected for the studies understand the structural similarities of PIG-L and GPI12p proteins.

Table 3.4: The accession numbers, type and strains of the Histone deacetylase superfamily proteins of archaeal kingdom that were used for the generation of the phylogenetic tree

<table>
<thead>
<tr>
<th>Accession number</th>
<th>Deacetylase type</th>
<th>Archea Strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABM81282.1</td>
<td>NAD-dependent protein Deacetylases, SIR2 family</td>
<td>Hyperthermus butylicus DSM 5456</td>
</tr>
<tr>
<td>ADC65634.1</td>
<td>histone deacetylase superfamily</td>
<td>Ferroglobus placidus DSM 10642</td>
</tr>
<tr>
<td>ZP_09946761.1</td>
<td>histone deacetylase</td>
<td>Halobiforma lacisalsi AJ5</td>
</tr>
<tr>
<td>Accession</td>
<td>Description</td>
<td>Organism</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>YP_004763380.1</td>
<td>acetylpolyamine amidohydrolase, histone-deacetyl superfamily</td>
<td>Thermococcus sp. 4557</td>
</tr>
<tr>
<td>YP_002959685.1</td>
<td>NAD-dependent deacetylase</td>
<td>Thermococcus gammatolerans</td>
</tr>
<tr>
<td>ADJ16326.1</td>
<td>histone deacetylase superfamily protein</td>
<td>Halalkalicoccus jeotgali B3</td>
</tr>
<tr>
<td>YP_003435909.1</td>
<td>histone deacetylase superfamily</td>
<td>Ferroglobus placidus DSM 10642</td>
</tr>
<tr>
<td>YP_003404410.1</td>
<td>polysaccharide deacetylase</td>
<td>Haloterrigena turkmenica</td>
</tr>
<tr>
<td>ACS33821.1</td>
<td>NAD dependent protein deacetylase, Sir2 family (npdA)</td>
<td>Thermococcus gammatolerans</td>
</tr>
<tr>
<td>ZP_08561746.1</td>
<td>histone deacetylase superfamily protein</td>
<td>Halorhabdus tiamatea SARL4B</td>
</tr>
<tr>
<td>ZP_06387924.1</td>
<td>acetoin utilization protein</td>
<td>Sulfolobus solfataricus 98/2</td>
</tr>
<tr>
<td>ZP_08045835.1</td>
<td>polysaccharide deacetylase</td>
<td>Haladaptatus pauchhalophilus</td>
</tr>
<tr>
<td>CAJ36490.1</td>
<td>putative acetoin utilization (histone deacetylase superfamily)</td>
<td>Methanocella arvoryzae MRE50</td>
</tr>
<tr>
<td>YP_003815998.1</td>
<td>Polysaccharide deacetylase</td>
<td>Acidilobus saccharovorans 345-15</td>
</tr>
<tr>
<td>ADG90809.1</td>
<td>acetylornithine deacetylase</td>
<td>Thermosphaera aggregans</td>
</tr>
<tr>
<td>ADC66595.1</td>
<td>histone deacetylase superfamily</td>
<td>Ferroglobus placidus DSM 10642</td>
</tr>
<tr>
<td>EHF10090.1</td>
<td>histone deacetylase superfamily</td>
<td>Methanolinea tarda NOBI-1</td>
</tr>
<tr>
<td>AFD00212.1</td>
<td>putative polysaccharide deacetylase</td>
<td>Methanocella cion radii HZ254</td>
</tr>
<tr>
<td>YP_004797323.1</td>
<td>acetoin utilization protein/histone deacetylase superfamily</td>
<td>Haloarcula hispanica ATCC 33960</td>
</tr>
<tr>
<td>ZP_09946342.1</td>
<td>polysaccharide deacetylase</td>
<td>Halobiforma lacisalsi</td>
</tr>
</tbody>
</table>
Table 3.5: The accession numbers, type and strains of the Histone deacetylase superfamily proteins of bacteria kingdom that were used for the generation of the phylogenetic tree

<table>
<thead>
<tr>
<th>Accession no.</th>
<th>Deacetylase type &amp; Archea Strain</th>
<th>Accession no.</th>
<th>Deacetylase type &amp; Archea Strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>YP_001212583.1</td>
<td>Deacetyl - Pelotomaculum thermopro.</td>
<td>EHY76958.1</td>
<td>h-d Pseudomonas tutzeri</td>
</tr>
<tr>
<td>ABC33191.1</td>
<td>h-d &amp; AUP Hahella chejuensis</td>
<td>EH63526.1</td>
<td>H-d Ectothiorhodospira sp.</td>
</tr>
<tr>
<td>ZP_02928271.1</td>
<td>SIR2 Verrucomicrobiutum spinosum</td>
<td>ZP_09755579.1</td>
<td>h-d Alishewanella jeotgali</td>
</tr>
<tr>
<td>ZP_02733279.1</td>
<td>h-d &amp; AUP Gemmata obscuriglobus</td>
<td>ZP_09627930.1</td>
<td>h-d Cupriavidus basilensis</td>
</tr>
<tr>
<td>ZP_00394995.1</td>
<td>h-d &amp; AUP Bacillus anthracis</td>
<td>YP_001630507.1</td>
<td>unknown Bordetella petrii</td>
</tr>
<tr>
<td>ZP_01627512.1</td>
<td>h-d &amp; AUP proteobacterium sp.</td>
<td>AEJ97012.1</td>
<td>put-d Klebsiella pneumonia</td>
</tr>
<tr>
<td>CAJ54674.1</td>
<td>h-d, AUP Lawsonia intracellularis</td>
<td>AEI65249.1</td>
<td>h-d Myxococcus fulvus</td>
</tr>
<tr>
<td>CBK85281.1</td>
<td>SIR2 Enterobacter cloaceae</td>
<td>AEF05127.1</td>
<td>H-d Alteromonas sp.</td>
</tr>
<tr>
<td>CBL34949.1</td>
<td>SIR2 Eubacterium siraeum</td>
<td>AEB44338.1</td>
<td>h-d Verrucosispora maris</td>
</tr>
<tr>
<td>CBL21367.1</td>
<td>SIR2 Ruminococcus sp.</td>
<td>ADV93770.1</td>
<td>pr-d Bacillus subtilis</td>
</tr>
<tr>
<td>Accession</td>
<td>Species</td>
<td>Strain</td>
<td>GenBank Accession</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------</td>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>CBL17622.1</td>
<td>SIR2 Ruminococcus champanellensis</td>
<td>ADO77617.1</td>
<td>p-d Halanaerobium praevalens</td>
</tr>
<tr>
<td>CBL13451.1</td>
<td>SIR2 Roseburia intestinalis</td>
<td>ADL07615.1</td>
<td>h-d Thermosediminibacter oceani</td>
</tr>
<tr>
<td>CK96161.1</td>
<td>SIR2 Eubacterium siraeum</td>
<td>ABM19578.1</td>
<td>h-d Marinobacter aquaeolei</td>
</tr>
<tr>
<td>CK95141.1</td>
<td>SIR2 Eubacterium rectale</td>
<td>ZP_09540222.1</td>
<td>H-d Nesterenkonia sp.</td>
</tr>
<tr>
<td>CK63003.1</td>
<td>SIR2 Alistipes shahii</td>
<td>EHP39604.1</td>
<td>h-d Cupriavidus basilensis</td>
</tr>
<tr>
<td>CK81290.1</td>
<td>SIR2 Coprococcus catus</td>
<td>EDZ63924.1</td>
<td>h-d beta proteobacterium</td>
</tr>
<tr>
<td>CK71407.1</td>
<td>SIR2 Bifidobacterium longum</td>
<td>EHM03968.1</td>
<td>unknown Bacillus amyloliquefaciens</td>
</tr>
<tr>
<td>CAM75635.1</td>
<td>SIR2 Magnetospirillum gryphiswaldense</td>
<td>AEV88021.1</td>
<td>h-d Actinoplanes sp.</td>
</tr>
<tr>
<td>ABD85719.1</td>
<td>SIR2 Rhodopseudomonas palustris</td>
<td>EHK83570.1</td>
<td>AUP Saccharomonospora azurea</td>
</tr>
<tr>
<td>ABB07643.1</td>
<td>SIR2 Burkholderia sp.</td>
<td>EHK62651.1</td>
<td>H-d Halomonas sp.</td>
</tr>
<tr>
<td>AAM24934.1</td>
<td>SIR2 Thermoanaerobac. Tengcongensis</td>
<td>EBJ05464.1</td>
<td>h-d Marinobacter manganoxydans</td>
</tr>
<tr>
<td>ZP_00393309.1</td>
<td>SIR2 Bacillus anthracis</td>
<td>ZP_09287416.1</td>
<td>H-d Halomonas sp.</td>
</tr>
<tr>
<td>ZP_04414274.1</td>
<td>SIR2 Vibrio cholera</td>
<td>ZP_09266709.1</td>
<td>h-d Leptospira weillii</td>
</tr>
<tr>
<td>ZP_00121115.1</td>
<td>SIR2 Bifidobacterium longum</td>
<td>ZP_09130213.1</td>
<td>P-d Desulfovibrio sp.</td>
</tr>
<tr>
<td>ZP_00051077.1</td>
<td>SIR2 Magnetospirillum</td>
<td>AEK40950.1</td>
<td>AUC mycolatopsis mediterranei</td>
</tr>
<tr>
<td>Accession</td>
<td>Name</td>
<td>Protein ID</td>
<td>Taxonomy</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------------------</td>
<td>------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>CBK99043.1</td>
<td>SIR2 Faecalibacterium prausnitzii</td>
<td>EAV47323.1</td>
<td>H-d Methylophilales bacterium</td>
</tr>
<tr>
<td>CBL31174.1</td>
<td>SIR2 Enterococcus sp.</td>
<td>EAS84856.1</td>
<td>H-d Candidatus Pelagibacter</td>
</tr>
<tr>
<td>CBL23557.1</td>
<td>SIR2 Ruminococcus obeum</td>
<td>EAR18236.1</td>
<td>put-H-d Synechococcus sp.</td>
</tr>
<tr>
<td>BAB98037.1</td>
<td>SIR2 Corynebacterium glutamicum</td>
<td>EAQ32203.1</td>
<td>H-d Idiomarina báltica</td>
</tr>
<tr>
<td>ZP_09918877.1</td>
<td>AUP Lentibacillus sp.</td>
<td>EAQ08202.1</td>
<td>H-d Loktanella vestfoldensis</td>
</tr>
<tr>
<td>ZP_09910691.1</td>
<td>h-d Sphingobium yanoikuyae</td>
<td>EAQ03894.1</td>
<td>H-d Oceanicola batsensis</td>
</tr>
<tr>
<td>ZP_09901277.1</td>
<td>h-d Methylomicrobium album</td>
<td>YP_003279482.1</td>
<td>H-d Comamonas testosteroni</td>
</tr>
<tr>
<td>EHK58917.1</td>
<td>h-d Mesorhizobium alhagi</td>
<td>ZP_09101230.1</td>
<td>H-d Desulfitomaculum gibsoniae</td>
</tr>
<tr>
<td>EEB72501.1</td>
<td>H-d Ruegeria sp.</td>
<td>ZP_09091397.1</td>
<td>H-d Mesorhizobium amorphae</td>
</tr>
<tr>
<td>AFH92707.1</td>
<td>h-d Providencia</td>
<td>ZP_08863119.1</td>
<td>H-d Ruegeria sp.</td>
</tr>
<tr>
<td>EIE16042.1</td>
<td>h-d Burkholderia sp.</td>
<td>ZP_08872456.1</td>
<td>H-d Verminephrobac. Aporrectodeae</td>
</tr>
<tr>
<td>ZP_09995399.1</td>
<td>h-d Acidithiobacillus Thi oxidants</td>
<td>EEF79185.1</td>
<td>H-d Methylphaga thiooxydans</td>
</tr>
<tr>
<td>ZP_09980618.1</td>
<td>h-d Mycobacterium xenopi</td>
<td>YP_003701553.1</td>
<td>H-d Syntrophothermus lipocalidus</td>
</tr>
<tr>
<td>AEX99797.1</td>
<td>H-d Oceanimonas sp.</td>
<td>ZP_08785184.1</td>
<td>AUP Ornithinibacillus scapharcae</td>
</tr>
<tr>
<td>CCE95060.1</td>
<td>h-d Sinorhizobium fredii</td>
<td>EGG61470.1</td>
<td>H-d Acidithiobacillus sp.</td>
</tr>
<tr>
<td>YP_001533392.1</td>
<td>h-d Dinoroseobacter shibae</td>
<td>ZP_06380062.1</td>
<td>H-d Arthrospira platensis</td>
</tr>
<tr>
<td>Accession</td>
<td>Species/Genus</td>
<td>Protein Accession</td>
<td>Name</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------</td>
<td>-------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>YP_003825238.1</td>
<td>h-d Thermosediminibacter oceani</td>
<td>ZP_01447469.1</td>
<td>H-d alphaproteobacterium</td>
</tr>
<tr>
<td>ADI00988.1</td>
<td>h-d Syntrophothermus lipocalidus</td>
<td>EGB97625.1</td>
<td>H-d Pseudomonas sp</td>
</tr>
<tr>
<td>CBA09794.1</td>
<td>h-d Neisseria meningitides</td>
<td>EHG06103.1</td>
<td>H-d Desulfitomaculum gibsoniae</td>
</tr>
<tr>
<td>EDZ46018.1</td>
<td>h-d Rhodobacterales bacterium</td>
<td>YP_002634430.1</td>
<td>put-AUP Staphylococcus carnosus</td>
</tr>
<tr>
<td>EDY93079.1</td>
<td>h-d Octadecabacter antarcticus</td>
<td>CBA04428.1</td>
<td>H-d Neisseria meningitides</td>
</tr>
<tr>
<td>YP_001341751.1</td>
<td>h-d Marinomonas sp.</td>
<td>ZP_08666172.1</td>
<td>H-d Paracoccus sp. TRP</td>
</tr>
<tr>
<td>ZP_09850211.1</td>
<td>AUP Brachybacterium squillarum</td>
<td>ZP_08550398.1</td>
<td>H-d Salinisphaera shabanensis</td>
</tr>
<tr>
<td>AEO49151.1</td>
<td>h-d Rhodospirillum rubrum</td>
<td>ZP_05586637.1</td>
<td>H-d Burkholderia thailandensis</td>
</tr>
<tr>
<td>YP_004277994.1</td>
<td>h-d Agrobacterium</td>
<td>ZP_08405009.1</td>
<td>H-d Hylemonella gracilis</td>
</tr>
<tr>
<td>EEE45936.1</td>
<td>h-d Labrenzia alexandrii</td>
<td>EGH98994.1</td>
<td>H-d Pseudomonas syringae</td>
</tr>
<tr>
<td>EDX90048.1</td>
<td>H-d Alcanivorax sp.</td>
<td>ZP_01049560.2</td>
<td>deacetylase Dokdonia donghaensis</td>
</tr>
<tr>
<td>ZP_09848981.1</td>
<td>h-d Serinicoccus profundii</td>
<td>EAQ39532.2</td>
<td>deacetylase Dokdonia donghaensis</td>
</tr>
<tr>
<td>ZP_01059515.1</td>
<td>h-d Leeuwenhoekiella blandensis</td>
<td>EFU26028.1</td>
<td>H-d Staphylococcus aureus</td>
</tr>
</tbody>
</table>
Table 3.6: The accession numbers, type and strains of the Histone deacetylases superfamily proteins of eukaryotes kingdom that were used for the generation of the phylogenetic tree

<table>
<thead>
<tr>
<th>Accession no.</th>
<th>Deacetylase type &amp; Archea Strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAY71730.1</td>
<td>Sir2 Komagataella pastoris</td>
</tr>
<tr>
<td>AAD40129.1</td>
<td>h-d Arabidopsis thaliana</td>
</tr>
<tr>
<td>XP_003194462.1</td>
<td>Sir2 Cryptococcus gattii</td>
</tr>
<tr>
<td>ADD19904.1</td>
<td>h-d Glossina morsitans</td>
</tr>
<tr>
<td>NP_001518.3</td>
<td>h-d Homo sapiens</td>
</tr>
<tr>
<td>NP_077038.2</td>
<td>h-d Mus musculus</td>
</tr>
<tr>
<td>NP_445900.1</td>
<td>h-d Rattus norvegicus</td>
</tr>
<tr>
<td>NP_989487.1</td>
<td>h-d Gallus gallus</td>
</tr>
<tr>
<td>AAH54208.1</td>
<td>h-d Xenopus laevis</td>
</tr>
<tr>
<td>XP_002291445.1</td>
<td>h-d Thalassiosira pseudonana</td>
</tr>
<tr>
<td>ABN81195.1</td>
<td>h-d Schistosoma mansoni</td>
</tr>
<tr>
<td>NP_001105402.1</td>
<td>h-d Zea mays</td>
</tr>
<tr>
<td>NP_200915.2</td>
<td>h-d Arabidopsis thaliana</td>
</tr>
<tr>
<td>NP_956069.1</td>
<td>h-d Danio rerio</td>
</tr>
<tr>
<td>EEF06319.1</td>
<td>h-d Populus trichocarpa</td>
</tr>
<tr>
<td>EDP01149.1</td>
<td>h-d Chlamydomonas reinhardtii</td>
</tr>
<tr>
<td>AAP95014.1</td>
<td>put-h-d Cochliobolus carbonum</td>
</tr>
<tr>
<td>AAP95013.1</td>
<td>put-h-d Emericella nidulans</td>
</tr>
<tr>
<td>AAK11184.1</td>
<td>h-d Rattus norvegicus</td>
</tr>
<tr>
<td>XP_001862395.1</td>
<td>h-d Culex quinquefasciatus</td>
</tr>
<tr>
<td>XP_002287324.1</td>
<td>h-d Thalassiosira pseudonana</td>
</tr>
<tr>
<td>ACD50313.1</td>
<td>h-d Hordeum vulgare</td>
</tr>
<tr>
<td>AFH32163.1</td>
<td>h-d Macaca mulatta</td>
</tr>
<tr>
<td>XP_003190418.1</td>
<td>h-d Aspergillus oryzae</td>
</tr>
<tr>
<td>NP_001125738.1</td>
<td>h-d Pongo abelii</td>
</tr>
<tr>
<td>NP_001083811.1</td>
<td>h-d Xenopus laevis</td>
</tr>
<tr>
<td>NP_001032521.1</td>
<td>h-d Bos Taurus</td>
</tr>
<tr>
<td>XP_844962.1</td>
<td>h-d Trypanosoma brucei</td>
</tr>
<tr>
<td>Accession</td>
<td>Species</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>XP_647498.1</td>
<td>h-d Dictyostelium discoideum</td>
</tr>
<tr>
<td>CCA72749.1</td>
<td>h-d Piriformospora indica</td>
</tr>
<tr>
<td>EGE86484.1</td>
<td>h-d Ajellomyces dermatitidis</td>
</tr>
<tr>
<td>EDV24263.1</td>
<td>unknown Trichoplax adhaerens</td>
</tr>
<tr>
<td>NP_001017414.1</td>
<td>SIR2 Gallus gallus</td>
</tr>
<tr>
<td>NP_001016883.1</td>
<td>h-d Xenopus tropicalis</td>
</tr>
<tr>
<td>NP_999711.1</td>
<td>h-d Strongylocentrotus purpuratus</td>
</tr>
<tr>
<td>EIE26198.1</td>
<td>h-d Coccomyxa subellipsoidea</td>
</tr>
<tr>
<td>CAL52346.2</td>
<td>h-d Ostreococcus tauri</td>
</tr>
<tr>
<td>EIE87309.1</td>
<td>unknown Rhizopus oryzae</td>
</tr>
<tr>
<td>XP_003705096.1</td>
<td>h-d Megachile rotundata</td>
</tr>
<tr>
<td>AAF52023.1</td>
<td>h-d Drosophila melanogaster</td>
</tr>
</tbody>
</table>

Formerly selected 17 sequences along with this Archea Histone deacetylase superfamily were used to generate a radial Phylogram better to understand the relationship of the PIG-L and GPI12p proteins by Archea Histone deacetylase superfamily proteins. These results are as follows: This method was also carried out for bacterial and eukaryote Histone deacetylase superfamily proteins in a similar fashion.
Fig. 3.11: A radial phylogram generated using MEGA v4.0 using Archea Histone deacetylase superfamily proteins with sample proteins (PIG-L and GPI12p proteins)
Fig 3.12: A radial Phylogram generated using MEGA v4.0 using Bacterial Histone deacetylase superfamily proteins with sample proteins (PIG-L and GPI12p proteins)
Fig 3.13: A radial Phylogram generated using MEGA v4.0 using Eukaryote Histone deacetylase superfamily proteins with sample proteins (PIG-L and GPI12p proteins)
3.9 DISCUSSION

Using the Maximum Parsimony method (MP) (1) inferred the history of evolution. Present Trees each 1 out of 4 most parsimonious trees (length = 1853) is shown. The consistency index is 0.608743 (0.608743), the retention index is 0.585714 (0.585714), and the composite index is 0.356549 (0.356549) for all sites and parsimony-informative sites (in parentheses). The MP tree was obtained using the Close-Neighbour-Interchange algorithm with search level 2 (1, 2). Phylogenetic analyzes were conducted in MEGA4 [144]. From the above figures of radial phylograms, it is clearly evident that the Histone deacetylase superfamily proteins have high degree of similarity with sample proteins (PIG-L and GPI12p proteins), which share several chunks of conserved sequences. Therefore, it can be readily inferred that these PIG-L and GPI12p proteins are greatly related to the deacetylase bacterial proteins rather than the archaea or eukaryotes as previously envisioned by many researchers.

Already some researchers by means, it was confirmed by the presence of GPI-anchored proteins in a strain of the genus Sulfolobus, which has closely related to eukaryotes. Our result supported by Eisenberg's group, and concluded that protein modification with GPI absent in all eubacterial and three Archea bacterial species analyzed whereas four Archea bacterial genomes has appeared the ability of GPI binding. GPI-anchor modification present in a certain group of Archaea bacteria as well as all Eukaryotes.