Chapter 4
Results and Discussion

"You do not know what you will find, you may set out to find one thing and end up by discovering something entirely different."

-Alexander Fleming

4.1 Characterization of metabolic pathway genes

4.1.1 Major metabolic pathways

All biochemical reactions occurring within the cells of an organism are considered as metabolism. Bacterial metabolism mainly focuses on the chemical diversity of substrate oxidations as well as dissimilation reactions (reactions through which substrates are broken down) (Peter, 1996). These reactions help the bacteria to generate energy. Moreover, uptake and utilization of several different organic and inorganic nutrients (assimilation reaction) for the growth and maintenance of microbes are also a part of bacterial metabolism. Thus, a series of exergonic and endergonic reactions catalyzed by integrated enzyme systems ultimately result in self-replication of cells (Peter, 1996). The capacity of bacterial cells to grow and function in proper chemical milieu (bacterial culture medium) along with the chemical changes resulting in bacterial transformation fall under the scope of bacterial metabolism. There are several metabolic pathways in each organism involving different proteins, enzymes and operons. Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/) is one of the major databases dealing with genomes, biological pathways, chemical substances, diseases and drugs (https://en.wikipedia.org/wiki/KEGG). This database has been used mainly in bioinformatics and metabolomics study.

We have also used this database for selecting the major metabolic pathways among considered actinobacteria strains. Ultimately we chose 8 major

metabolic pathways. They are – (1) Carbohydrate metabolism pathway (CM), (2) Amino acid metabolism (AM), (3) Energy metabolism (EM), (4) Lipid metabolism (LM), (5) Nucleotide metabolism (NM), (6) Co-factor and vitamin metabolism (CVM), (7) Terpenoids and polyketides metabolism (TPM) and (8) Xenobiotics biodegradation and metabolism (XBM). CM included Amino sugar and nucleotide sugar metabolism, Ascorbate and aldarate metabolism, Butanoate metabolism, C5-Branched dibasic acid metabolism, Citrate cycle (TCA cycle), Fructose and mannose metabolism, Galactose metabolism, Glycolysis / Gluconeogenesis, Glyoxylate and dicarboxylate metabolism, Inositol phosphate metabolism, Pentose and glucuronateinter-conversions, Pentose phosphate pathway, Propanoate metabolism, Pyruvate metabolism, Starch and sucrose metabolism. AM included Alanine, aspartate and glutamate metabolism, Arginine and proline metabolism, Arginine biosynthesis, Cysteine and methionine metabolism, Glycine, serine and threonine metabolism, Histidine metabolism, Lysine biosynthesis, Lysine degradation, Phenylalanine metabolism, Phenylalanine, tyrosine and tryptophan biosynthesis, Tryptophan metabolism, Tyrosine metabolism, Valine, leucine and isoleucine biosynthesis, Valine, leucine and isoleucine degradation. Carbon fixation pathways in prokaryotes, Methane metabolism, Nitrogen metabolism, Oxidative phosphorylation, Sulfur metabolism were included in EM. LM comprised of alpha-Linolenic acid metabolism, Arachidonic acid metabolism, Biosynthesis of unsaturated fatty acids, Ether lipid metabolism, Fatty acid biosynthesis, Fatty acid degradation, Glycerolipid metabolism, Glycerophospholipid metabolism, Linoleic acid metabolism, Primary bile acid biosynthesis, Sphingolipid metabolism, Steroid hormone biosynthesis and Synthesis and degradation of ketone bodies. NM included Purine metabolism and Pyrimidine metabolism. Biotin metabolism, Folate biosynthesis, Lipoic acid metabolism, Nicotinate and nicotinamide metabolism, One carbon pool by folate, Pantothenate and CoA biosynthesis, Porphyrin and chlorophyll metabolism, Retinol metabolism, Riboflavin metabolism, Thiamine metabolism, Ubiquinone and

other terpenoid-quinone biosynthesis and Vitamin B6 metabolism pathways were included in CVM. TPM composed of Biosynthesis of ansamycins, Biosynthesis of siderophore group nonribosomal peptides, Biosynthesis of type II polyketide products, Biosynthesis of vancomycin group antibiotics, Carotenoid biosynthesis, Geraniol degradation, Limonene and pinene degradation, Polyketide sugar unit biosynthesis, Sesquiterpenoid and triterpenoid biosynthesis, Terpenoid backbone biosynthesis, Tetracycline biosynthesis and Zeatin biosynthesis pathways.Aminobenzoate degradation, Atrazine degradation, Benzoate degradation, Bisphenol degradation, Caprolactam degradation, Chloroalkane and chloroalkene degradation, Chlorocyclohexane and chlorobenzene degradation, Dioxin degradation, Drug metabolism - cytochrome P450, Drug metabolism - other enzymes, Ethylbenzene degradation, Fluorobenzoate degradation, Metabolism of xenobiotics by cytochrome P450, Naphthalene degradation, Nitrotoluene degradation, Polycyclic aromatic hydrocarbon degradation, Steroid degradation, Styrene degradation, Toluene degradation and Xylene degradation were involved in XBM pathway.

4.1.2 Expression pattern analysis

Codon adaptation index (CAI) was calculated for estimating the expression pattern of the proteins (Wu et al. 2005) encoded by different metabolic pathway related genes. CAI has the capacity to directly measure the mRNA expression pattern (Roymondal et al. 2009, Wu et al. 2005, Eyre-Walker 1996, Coghlan and Wolfe 2000, Martn-Galiano et al. 2004, Dos et al. 2003) hence, depending upon the CAI values we have grouped the total protein coding genes of an organism among PHX, PMX and PLX category (as discussed in the Materials and Methods section). We checked out the distribution pattern of the investigated metabolic pathway related proteins. The results have been given as Fig 4.1. It is evident from the result that, most (76%) of the metabolic pathway related genes were under PHX category. Their distribution was limited to 1-2% in PLX category the rest being under PMX group. This is quite expected since; these genes were associated especially with the metabolism cum utilization of different nutrients and production of energy.

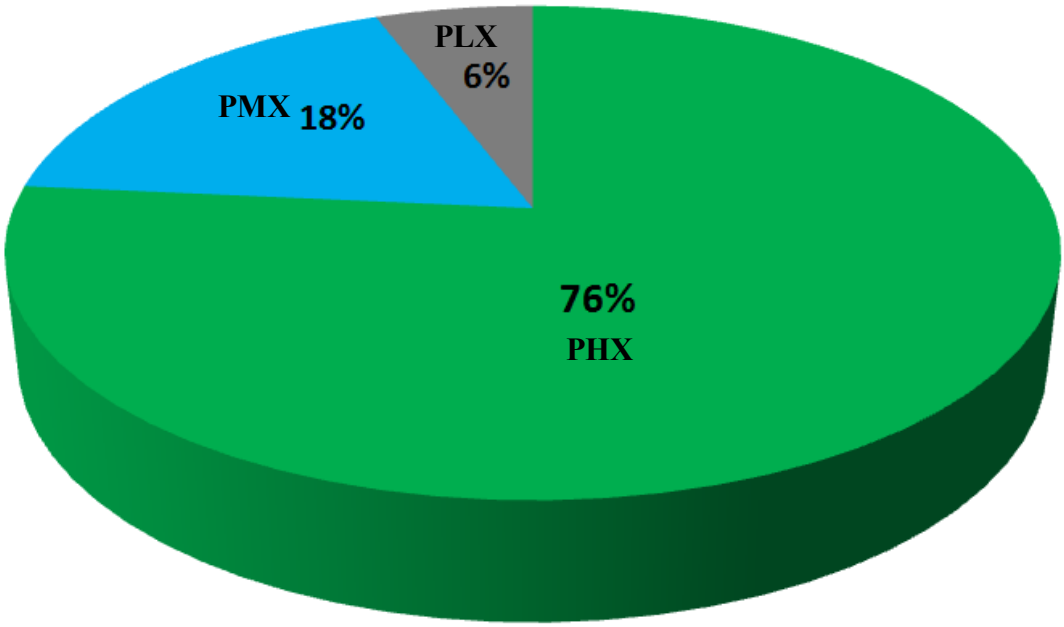


Fig 4.1 Pie chart showing the distribution of the metabolic pathway related genes among PHX, PMX and PLX categories. The values were calculated for each strain separately and then average value of all the considered metabolic pathways were taken for this plot generation.

This finding was also supported by Willenbrock and Ussery (2007). They showed the effect of position preference over genes associated with different metabolic pathway among prokaryotes. Position preference is basically a measurement strategy of DNA structure (Brukner et al. 1995). This concept actually originated from eukaryotes. Since DNA transcription is highly dependent on the bending pattern and flexibility of DNA, the structural pattern of DNA may reflect or influence the codon usage pattern also. Position preference is an idea to focus on the preference of a tri-nucleotide to be placed at a site where the DNA minor groove is either towards or away from the histone core of nucleosome (Satchwell et al. 1986). Higher inclination for nucleosome is reflected by high absolute position preference (Pedersen et al. 1998). On the contrary, lower position preference was assigned to those tri-nucleotides which tend to exclude nucleosomes. Unfortunately, this model was not suitable for prokaryotes since they lack nucleosome. However, both prokaryotes and eukaryotes have chromatin and similar level of DNA condensation level (more than 1000x).

Willenbrock and Ussery (2007) utilized these facts and modified the position preference model for prokaryotes. According to their hypothesis, position preference of a DNA can reflect how easily the DNA can be wrapped around the chromatin. It was revealed that, position preference was inversely proportional to the DNaseI sensitivity and directly proportional to the chromatic condensation preference (Willenbrock and Ussery 2007). Thus, the more is the preference value the less is the DNaseI sensitivity and more is the chromatin condensation preference at that particular part of DNA (Willenbrock and Ussery 2007). Conversely, the less is the preference value the more is the DNaseI sensitivity and less is the chromatin condensation preference of DNA. Ribosomal genes and other highly expressed (actively dividing) genes which were found to be with low position preference (Dlakić et al. 2005). They were more sensitive to DNaseI and were not much condensed. This actually helped them to access the RNA polymerase radially and make them easy to be transcribed. Thus, a correlation between the position preference and expression level of genes were established (Dlakić et al. 2005). Moreover, Willenbrock and Ussery (2007) also revealed that, metabolic pathway related genes were generally under low position preference category which implied their high expression and vitality in microbial genomics.

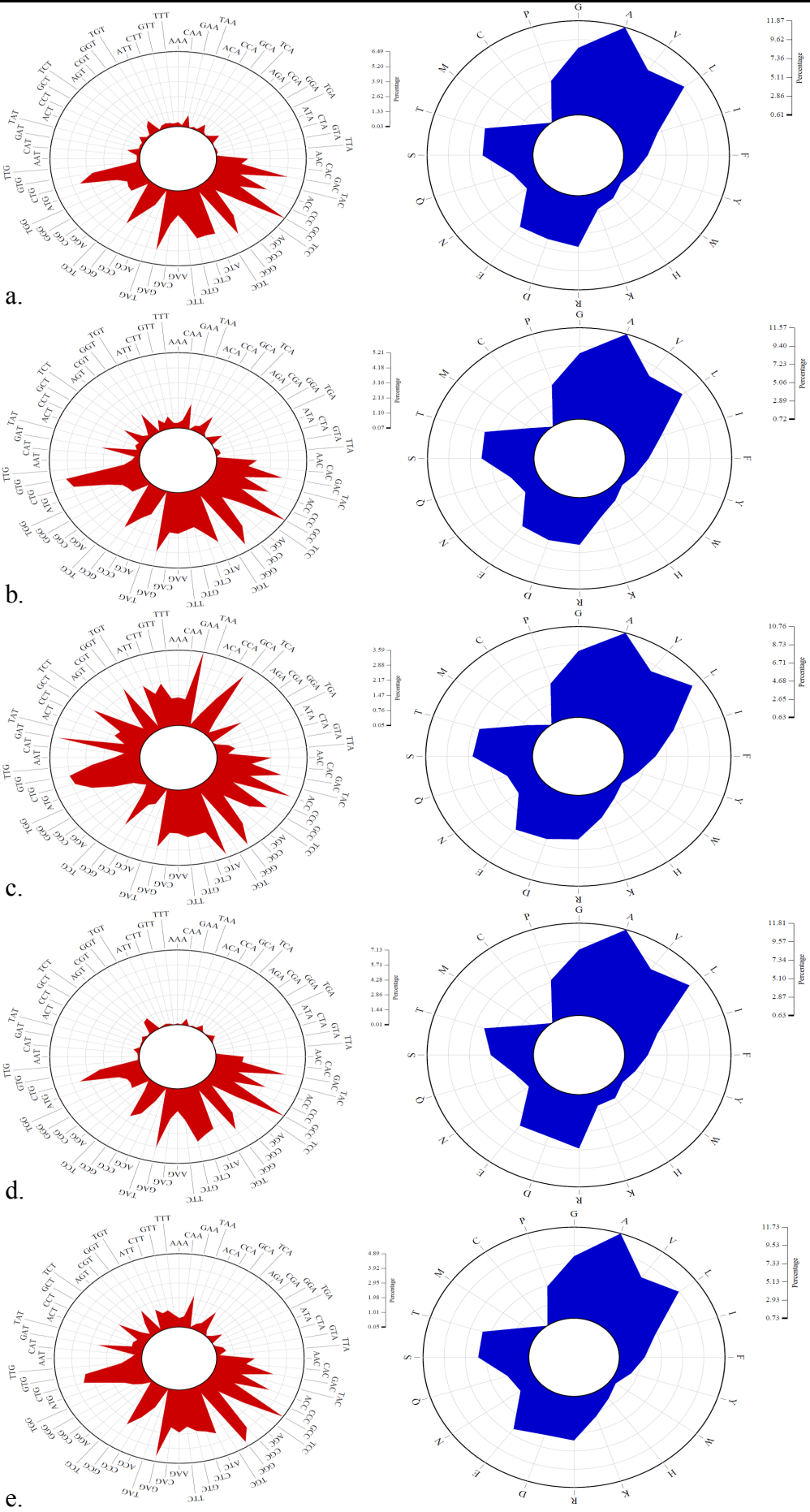
4.1.3 Effect of compositional constrain

Most of the actinobacterial genera are GC rich and tends to use the more GC3 rich codons among protein coding genes. Studies have also revealed that the preference of GC3 rich codons is higher among PHX rather than PLX (Roy et al. 2015). Correlating the CAI values with GC content may prove to be a good measure of compositional constrain among genes. Supplementary figure 1 showed the correlation coefficients among GC3 and CAI values of selected metabolic pathway related genes among investigated actinobacteria. The colour code has been given at the bottom of the figure. For AT rich strains, correlation between AT3 and CAI was considered. This result indicated a clear persistence of compositional constrain on these genes. Codons like GCC, GCG, CTG, CGC, GAC, GTG, GGC and ACC were some of the predominantly utilized codons in most of the cases (Fig 4.2 and 4.3) except *Tropheryma*

whipplei and *Gardnerella vaginalis*. This was expected, since these two genera are exception from rest of the actinoabcteria considering their genomic composition. They are low GC content actinobacteria (Das et al. 2006, Yeoman et al. 2010) and have negative correlation between CAI and GC3.

The positive correlation between CAI and GC3 among metabolic pathway related genes revealed that, preference for GC rich codons enhances with increased expression level. To further validate the result, we also did an amino acid usage analysis and found alanine, valine, leucine, glycine, arginine and aspartic acid to be most preferred amino acids among the metabolic pathway related proteins (Fig 4.2). Analysis on their synonymous codon usage pattern revealed that, while encoding these aforementioned amino acids, GC rich synonymous codons were preferred over AT rich ones. On contrary, *T. whipplei* and *G. vaginalis* showed the preference of AT rich synonymous codons while encoding the same amino acids. This finding further emphasizes the persistence of compositional constrain (be it GC constrain or AT constrain) over the considered metabolic pathway related genes among selected strains.

Wright (1990) suggested the correlation between GC3 and ENc can be an effective measure to study the codon usage variation among genes and genomes. ENc value generally ranges from 20-62 (Peden 1999). Heatmap (Supplementary Figure 2) based on the correlation coefficient between CAI and ENc revealed significant negative correlation between these two parameters indicating the fact that, ENc value decreases with increase in CAI value. This means, PHX genes have lower ENc values whereas PLX genes are with higher ENc values. Metabolic pathway related genes considered in this study posses be very low ENc value (ranges from 28-42). GC3 was correlated with ENc values and results have been given as Supplementary Figure 3. Significant negative correlation between ENc and GC3 values (significant positive correlation was found between GC3 and ENc of *T. whipplei* and *G. vaginalis*due to their AT compositional constrain) depicted strong codon usage bias among the considered genes (Comeron and Aguade, 1998; Roy et al. 2015) which further validates compositional


a.
b.
c.
d.
e.
Percentage

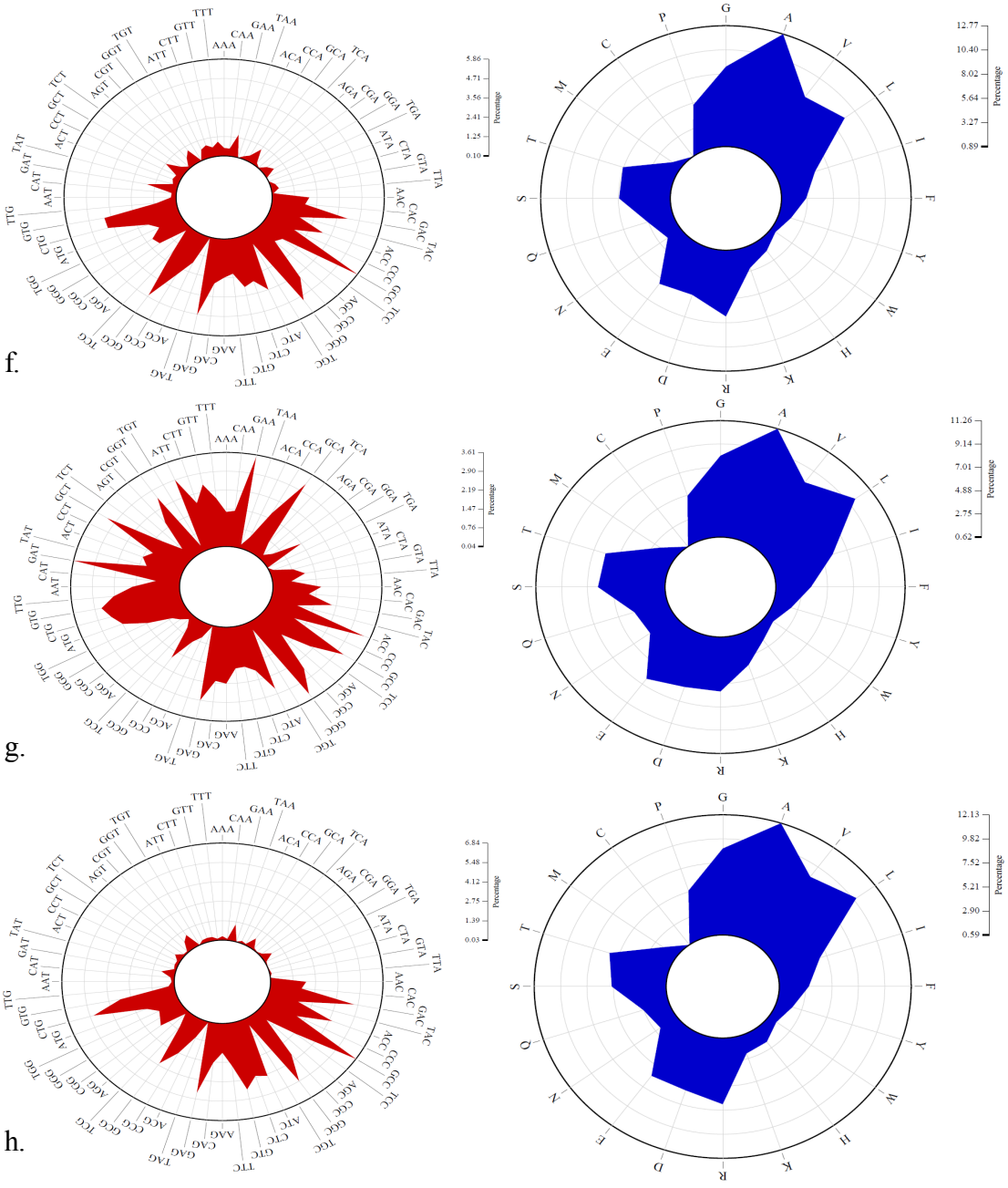


Fig 4.2 Rose plots for codon and amino acid usage analysis of 8 different metabolic pathways. Circles at the left side are for codon usage analysis (red) and on the right side are for amino acid usage (blue). Different metabolic pathways selected for this analysis are- (a) carbohydrate metabolism, (b) amino acid metabolism, (c) energy metabolism, (d) lipid metabolism, (e) nucleotide metabolism, (f) co-factors and vitamin metabolism, (g) terpenoids and polyketides metabolism and (h) xenobiotics biodegradation and metabolism. *Tropheryma whipplei* and *Gardnerella vaginalis* were not considered in this analysis.

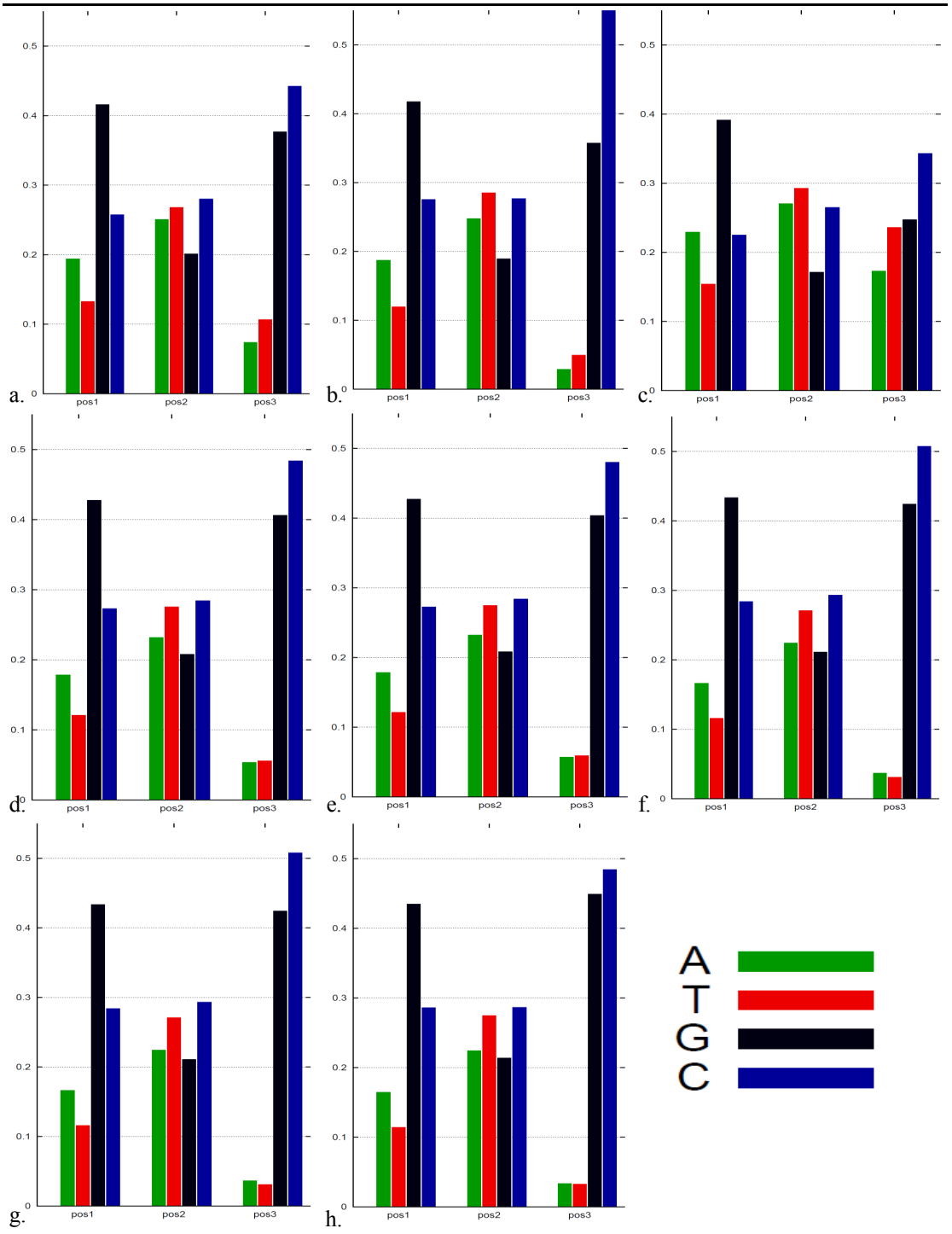


Fig 4.3 Bar diagram of position specific nucleotide usage of 8 different metabolic pathways. Different metabolic pathways selected for this analysis are- (a) carbohydrate metabolism, (b) amino acid metabolism, (c) energy metabolism, (d) lipid metabolism, (e) nucleotide metabolism, (f) co-factors and vitamin metabolism, (g) terpenoids and polyketides metabolism and (h) xenobiotics biodegradation and metabolism. *Tropheryma whipplei* and *Gardnerella vaginalis* were not considered in this analysis.

constrain as one of the major indices governing the codon and amino acid usage pattern of selected genes and their respective transcribed proteins.

4.1.4 Importance of optimal codons

Optimal codons are defined as the codons mostly preferred by the PHX genes (Peden 1999). Heatmap (Supplementary Figure 4) revealed a positive correlation between CAI and Fop among selected genes. This indicates, the preferential usage of optimal codons among the metabolic pathway related genes. Further, the positive correlation between Fop and GC3 also signifies that optimal codons are GC biased (except *T. whipplei* and *G. vaginalis*, their optimal codons were AT rich). GCC, CGG, GCG, GGG, CTG, GAC, CCC, CCG, CGC and GGC were most commonly used optimal codons among considered gene set of studied GC rich actinobacteria. On the contrary, ATT, TAT, CTT, CTA, GAT, CAT, GTT were some of the frequently used optimal codons among *T. whipplei* and *G. vaginalis.*

4.1.5 Effect of translational selection

A detailed analysis on the anticodon information along with tRNA copy number was done. This displayed a clear pattern revealing the fact that; majority of the optimal triplets (codons) was perfectly recognized by most of the abundant iso-accepting tRNAs (Roy et al. 2015). Hence, those codons were both optimal as well as translationaly optimal codons. GGC, CGC, CGG, GCC, CCC CAT, CTA, GTT, TAT, ATT were optimal as well as translationally optimal in *T. whipplei* and *G. vaginalis*. tAI calculation further probed the effect of translational selection over the considered gene set. Significant positive correlation between CAI and tAI along with tAI and Fop (Supplementary Figure 5, 6) among all investigated strains reinforced the effect of both natural selection and translation efficiency on their metabolic pathway related genes (Roy et al. 2015). The effect of tRNA population over genomic codon usage pattern was previously reported in some other organisms like *D. melanogaster* (Moriyama and Powell1997), *E. coli* (Ikemura 1981) and *C. elegans* (Duret2000) etc.

Thus, CAI, compositional constrain, frequency of optimal codons (Fop) and tRNA adaptation index (tAI) were found to be the major indices governing codon usage pattern of genes associated with 8 considered

metabolic pathways among selected actinobacteria. A combinatorial effect of both natural selection and translational efficacy was also evident from this study.

4.2. Evaluation of protein biosynthetic energy cost

Protein biosynthetic energy cost (EC) of potentially highly expressed proteins (PHXPs), potentially medially expressed proteins (PMXPs) and potentially lowly expressed proteins (PLXPs) among selected actinobacteria from different niche was evaluated and compared. This analysis exposed some very interesting result discussed below.

4.2.1 Statistical analysis between the CAI and EC

The correlation coefficient between CAI and EC of different habitat specific actinobacteria was determined. These two parameters were negatively correlated among animal-host-associated actinobacteria (R^2 = 0.756) (Fig 4.4). On the contrary, strong positive correlation between CAI and EC was found in extremophilic (R^2 = 0.955), plant-associated (R^2 = 0.801), aquatic (R^2 = 0.817) and soil dwelling (R^2 = 0.746) actinobacteria (Fig 4.4). Moreover, ANOVA test followed by F-test and t-test on the mean EC values of PHXPs, PMXPs, and PLXPs for each strain revealed considerable differences among them (at the $p < 0.01$ level of significance). From this statistical analysis it was clear that PHXPs of animal host associated actinobacteria consume less biosynthetic energy than PMXPs and PLXPs whereas in other niche associated actinobacteria the EC values of PHXPs were found to be more than PMXPs and PLXPs. All the statistical data are available in supplementary file 3. A list of PHXP, PMXPs and PLXPs is given as supplementary file 4.

4.2.2 Regulation of amino acid biosynthetic pathway

Hodgson (2000) carried out a detailed study on the regulation of metabolic pathway in *Streptomyces.* We also tried to get deep view knowledge on the amino acid biosynthetic pathways of investigated actinobacteria. The modular architecture of amino acid biosynthetic pathway was obtained from KEGG database (Fig 4.5). The total amino acid biosynthesis pathway was found to be divided into two parts-core and extension. Conversion of Glyceraldehyde-3-phosphate to pyruvate along with serine and glycine associated pathways were under core part. The extensions of aforementioned

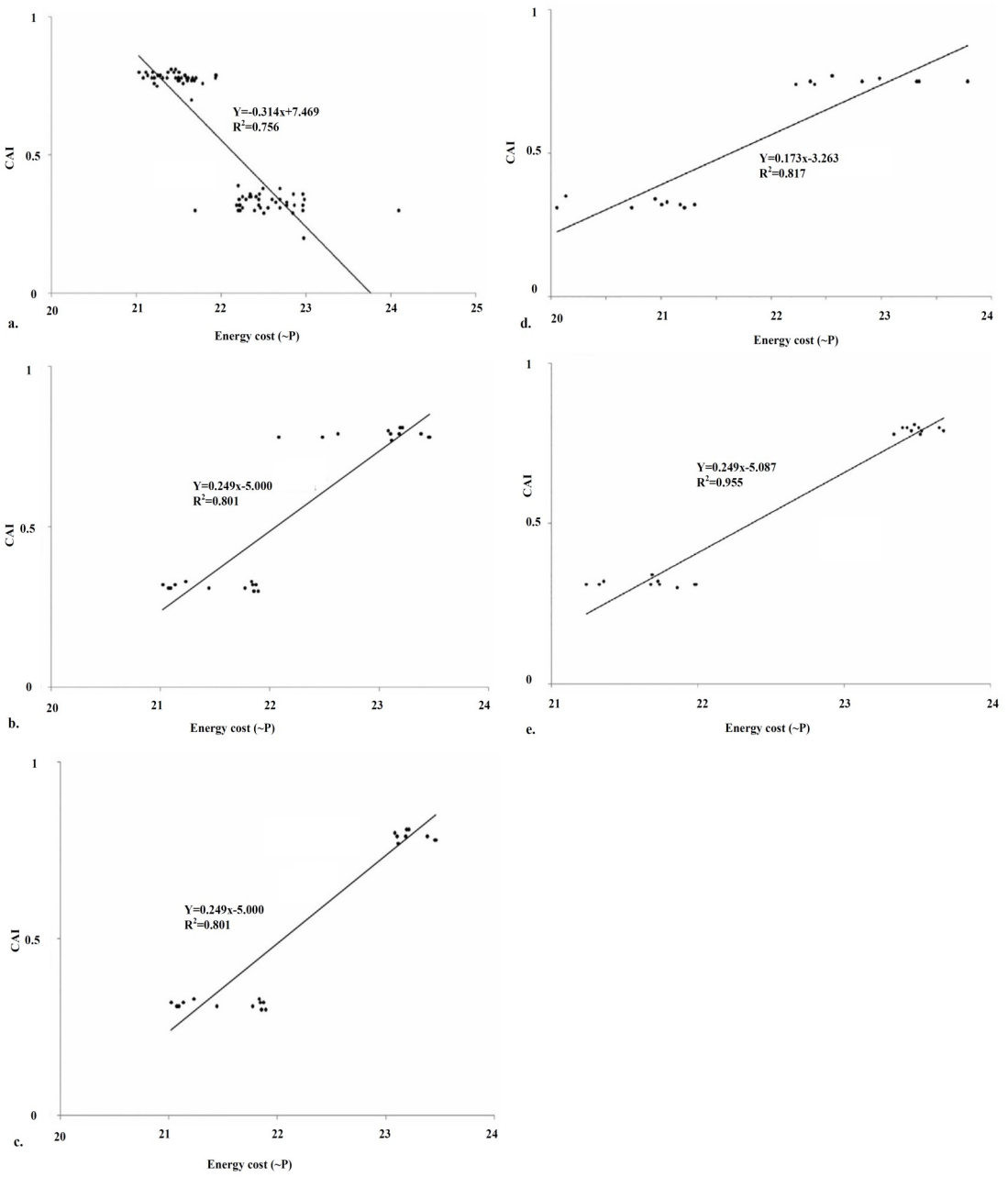


Fig 4.4 Correlation between CAI and Energy cost of selected Actinobacteria thriving in diverse niche (a) animal host associated, (b) plant host associated, (c) soil dwelling, (d) aquatic and (e) extremophilic actinobacteria. Only animal host associated actinobacteria showed negative correlation between CAI and EC whereas positive correlation was revealed among other niche specific actinobacteria.

pathways involved the reaction modules RM001, RM033, RM032, and RM002. RM001 and RM032 were found to be associated with 2-oxocarboxylic acid chain extension and carboxyl to amino conversion reactions, respectively. RM033 was involved in branched chain amino acid synthesis. RM002 was responsible for basic amino acid synthesis. This

extension is most divergent consisting of multiple lysine synthesis pathways along with multiple gene sets for arginine production. The aforesaid extensions were also responsible for essential amino acid biosynthesis. However, the most interesting fact is the feedback inhibition mechanism of aromatic amino acid biosynthesis pathway or shikimate pathway. A previous study on *Amycolatopsis methanolica* (a methylotrophic actinomycete) revealed the feedback mechanism of enzymes involved in shikimate pathway via modification of branch point intermediates (Herrmann and Weaver, 1999; Euverink 1995). Hence we may infer that, the

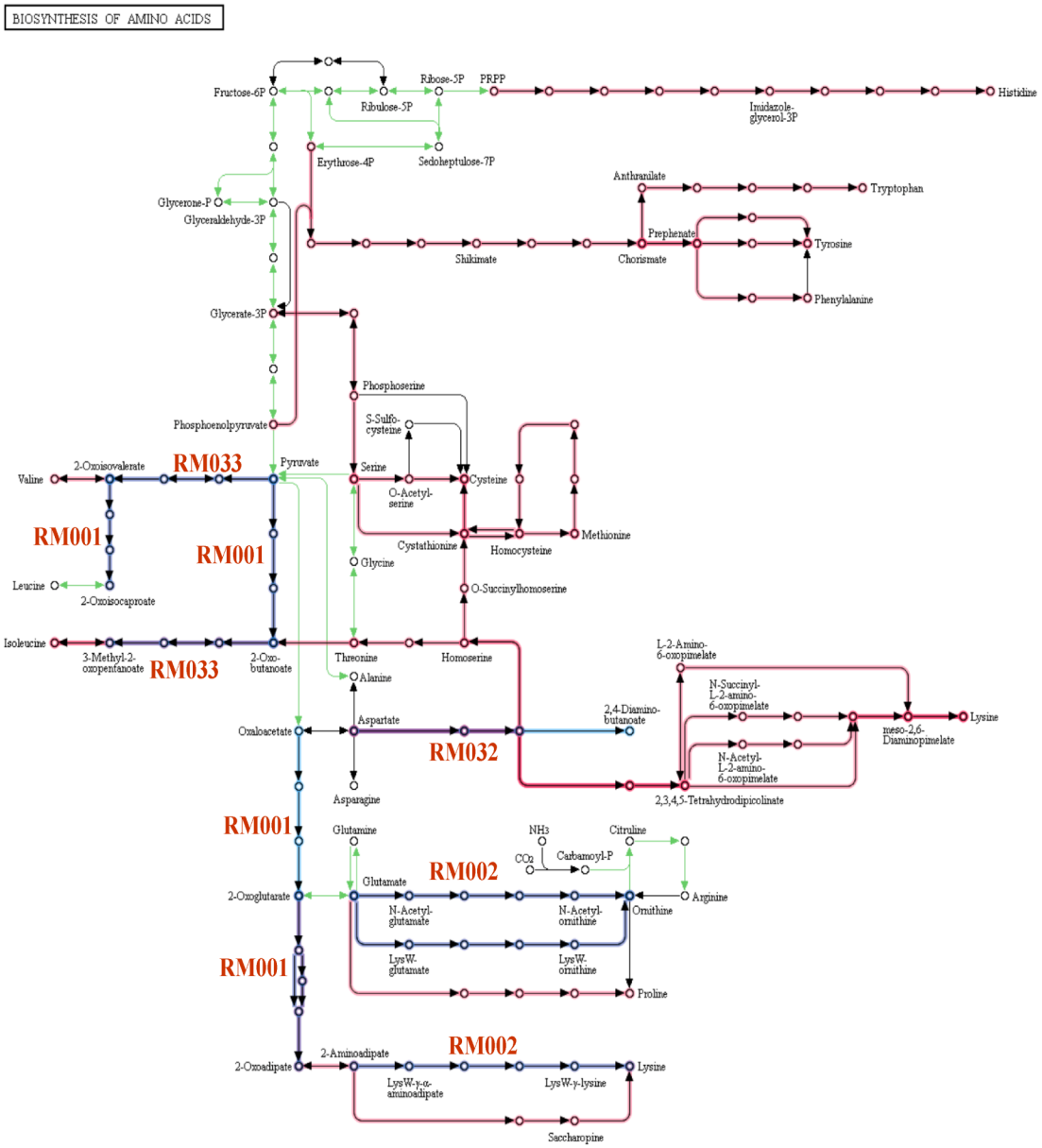


Fig 4.5 Amino acid biosynthesis pathway of Actinobacteria as depicted in KEGG database. Extensions are written in bold red color.

expression level of proteins involved in shikimate pathway is repressed when aromatic amino acids are adequate in environment.

4.2.3 Differential distribution of aromatic amino acids

Heatmap (Fig 4.6) illustrates the aromatic amino acid usage of PHXPs and PLXPs among the considered strains clearly revealed the recurrent exploitation of aromatic amino acids by the PHXPs of diverse niche specific actinobacteria except the animal host associated strains. This unusual finding may correlate with the niche adaptive capability of actinobacteria.

Animal host associated actinobacteria may obtain a substantial fraction of their amino acids from their respective host (Graf and Ruby 1998). Thus, they are in a nutrient rich environment and are not always dependent upon the amino acid biosynthesis. Host-derived amino acids are also important in proliferation and infection of pathogens. The animal partner provides a considerable proportion of amino acids along with other organic compounds to bacteria which are beneficial for both host and microbes. *Bifidobacterium* may behave as an auxotroph by obtaining certain amino acids from their host (Ferrario et al. 2015). Similarly a unidirectional flow of nutrients from host to pathogens is believed to support the infection cycle (Graf and Ruby 1998). *Chlamydia trachomatis* possess incomplete amino acid biosynthesis pathways and are thus dependent on the host for some essential amino acids like phenylalanine, tyrosin, tryptophan, glycine etc (Stephens et al. 1998). Since pathogens rely on their host for some essential amino acids including the aromatic (most costly) ones, natural selection may support the usage of energy economic amino acids in PHXPs. Furthermore, many of the PHX genes of animal-associated actinobacteria (both probiotic and pathogenic) contain signal peptides and were components of secretomes, which are essential for the host-microbe interactions (Roy et al. 2013). The proteins encoded by these genes were found to be using fewer aromatic amino acids than PLXPSs, lowering the overall energy cost of the PHXPs. This finding also confirms previous studies of (Heizer et al. 2006, Roy et al. 2013).

On the contrary, PHXPs of other niche associated actinobacteria (desert-, stone -, plant-, soil-, and water-) was found

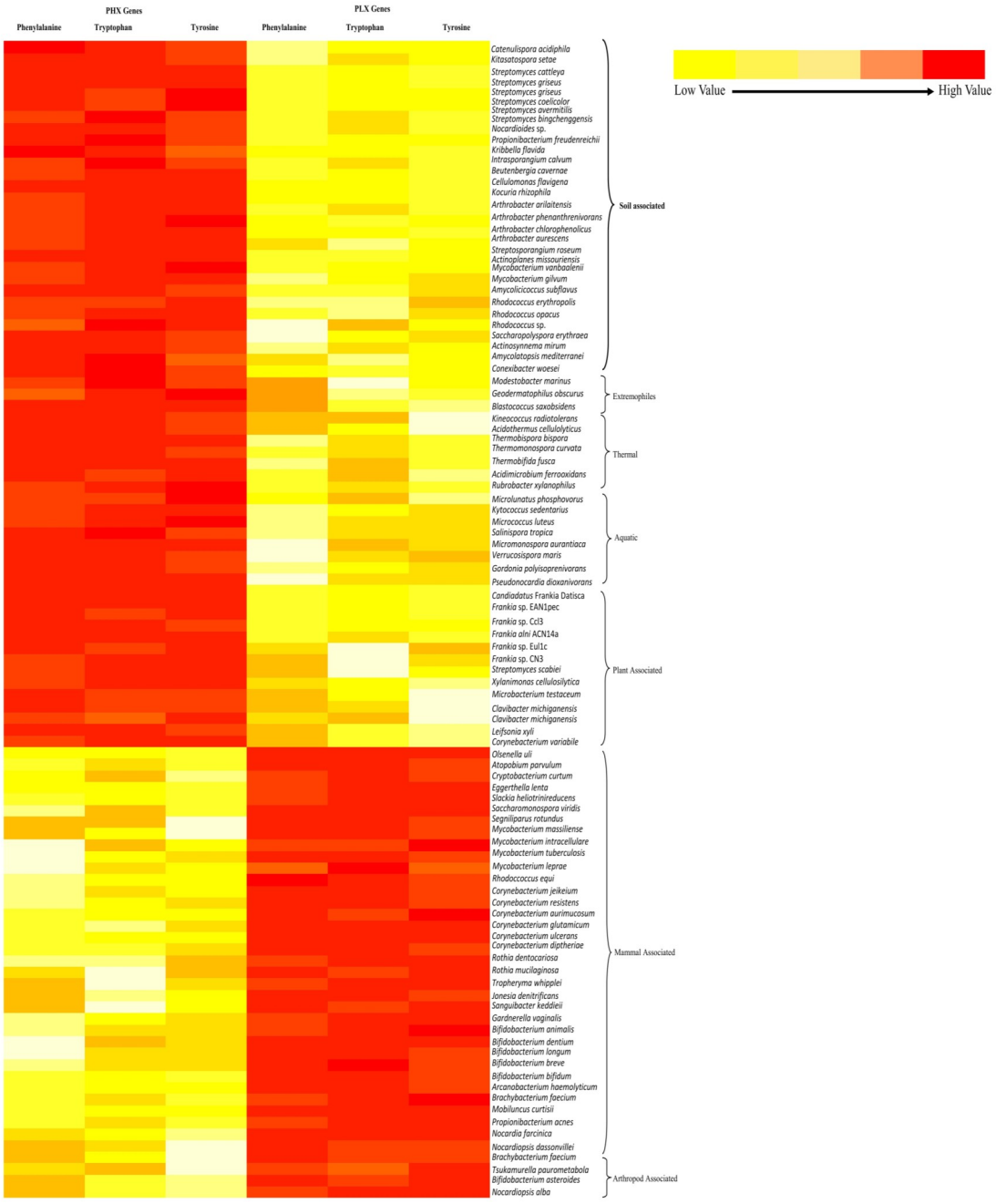


Fig 4.6 Heatmap on differential distribution pattern of costly aromatic amino acids in PHXPs and PLXPs of selected actinobacteria thriving in diverse niche.

to be using more aromatic amino acids. The PHXP set of actinobacteria thriving under harsh environmental conditions (stone surfaces, desert, hydrothermal vents, saline water, etc.) were found to contain shikimate pathway related proteins along with heat resistance proteins. Proteins involved in shikimate pathway were found to be using more aromatic amino acids increasing the overall EC of PHXP set. The products of this pathway also act as precursors of myco -sporines and mycosporines-like-amino

-acids (MAA) (Oren and Gunde-Cimerman 2007). These compounds are well known for their photo-protective and UV radiation resistant properties (Shick and Dunlap 2002) and thus play an imperative role in maintaining the healthy lifestyle of stone-dwelling and extremophilic actinobacteria. Other PHXPs of extremophilic actinobacteria were found to be involved in spore-formation, signal transduction, motility, cell division as well as metabolism and they utilize significant amount of aromatic amino acids. Moreover, a modest increase in thermo -stable phenylalanine and tyrosine usage in thermophiles and hyper-thermophiles has been reported previously (Oren and Gunde-Cimerman 2007). Thus, in the hyper-thermophiles and thermophiles aromatic amino acids may help the bacteria to thrive under extreme environmental circumstances. The π-orbitals (molecular orbital of the pi bond) of aromatic amino acids play a pivotal role in several ionic interactions providing better protein stabilization and thus may help in proper and stabilized protein-protein interactions acting as a survival strategy among extremophilic actinobacteria under ruthless situation (Goldstein 2007; Jaenicke and Bahm 1998; Kannan and Vishveshwara 2000).

In aquatic actinobacteria (like *Micrococcus, Micromonospora* etc.), UV radiation resistant proteins and tryptophan metabolism related proteins were mainly found under PHXP category. Interestingly those proteins were under PMXP category among *Cyanobacteria* (Supplementary File 5). Aquatic actinobacteria being predominantly present on the water surface are more exposed to UV radiation. UV radiation acts on both aromatic amino acids as well as dissolved oxygen matter (DOM) (Reitner et al. 2002). Generally, aquatic actinobacteria lack the UV-absorbing pigments and the below mentioned interactions further help the microbes to fight against the UV radiation. Tryptophan enhances the absorption rate of lower photosynthetically active radiation (PAR) and UV radiation via photo-coloring or humification of aquatic microorganisms (Shick and Dunlap 2002). Whereas in Cyanobacteria, UV absorbing pigments protects them from UV-radiation (Matsui et al. 2012).

Plant host associated microbes like *Frankia* (actinobacteria), *Rhozibium*

(proteobacteria) etc. produce plant hormones like indole acetic acid(IAA) and phenyl acetic acid (PAA) which are essential for plant growth (Prasad et al. 2000; Pedraza et al. 2004). The biosynthesis of IAA and PAA require aromatic amino acids for the amino transferase acitivity (Peret et al. 2008). In *Rhizobium*, proteins involved in IAA and PAA biosynthetic pathway were found under PMXP category whereas, in *Frankia*, those proteins were under PHXP category. Other PHXPs of *Frankia* were involved in secondary metabolite production, carbohydrate metabolism, signal transduction, and stress regulation. These proteins were found to be using phenylalanine, tyrosine, tryptophan, histidine, isoleucine, lysine, leucine, arginine, and valine more frequently than the rest of the amino acids. Majority of these amino acids are costly (Akashi and Gojobori 2002). Moreover, some host plants like alfalfa, soybean, and cowpea (host plants for *Rhizibium*) provide several amino acids like cysteine, arginine, isoleucine, tryptophan, valine, and histidine to their micro-symbionts (Randhawa and Hassani 2002). On the contrary, during nitrogen fixation, *Frankia* provides the nitrogen in the form of amides and amino acids to their hosts (Mus et al. 2016). This could be one of the reasons why the Rhizobeaceae maintain a low EC value in PHXPs and become auxotrophic for some amino acids. Heatmap (Fig 4.7) revealed higher usage of arginine, histidine, isoleucine, lysine, leucine, valine, tyrosnine, tryptophan, and phenylalanine in frankial PHXP rather than Rhizobiaceae.

Soil inhabiting *Streptomyces* are mainly important for antibiotic production however, they also produce phytohormones including gibberelin, auxins, cytokinins etc (Baca et al.1994; Subramaniam et al.2016; Hsu 2010). Proteins involved in antibiotic biosynthesis were found to be exploiting more aromatic amino acids however; the exact reason is still unknown. Moreover, a large set of Carbohydrate activating enzymes (CAZymes) were found in the PHXP set of *Streptomyces* as well as *Frankia*. These set of protein help the microbes to degrade complex polysaccharides into simple monosaccharides. Thus, they are also important in soil and plant host associated bacteria. These enzymes require aromatic amino acids at their active site for a stable interaction with their respective

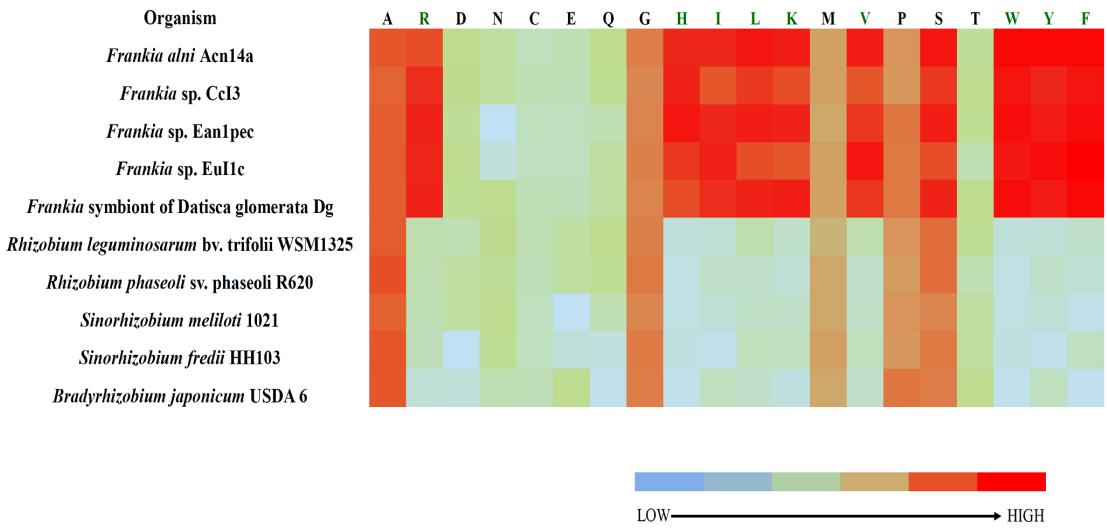


Fig 4.7 Heatmap on differential amino acid usage pattern among selected members of *Frankia* and Rhizobiaceae. Higher usage of arginine, histidine, isoleucine, lysine, leucine, valine, tyrosnine, tryptophan, and phenylalanine in frankial PHXP rather than Rhizobiaceae was clearly evident from this figure.

carbohydrate moiety (Zhang et al.2015; Tian et al. 2016). As a consequence the EC of PHXPs in soil and plant associated bacteria rises up.

The EC of PHXP, PMXP and PLXP of selected proteobacteria, firmicuites, cyanobacteria and yeasts were also calculated and compared with the results obtained from actinobacterial dataset. It was revealed that, in proteobacteria, firmicuites and cyanobacteria PHXPs were energy economic rather than PMXPs and PLXPs. Only the EC value of yeast PHXPs were more than their PMXPs and PLXPs (we could not do a statistical analysis due to shortage of data) (Supplementary File 6).

Thus, this analysis revealed a strange finding regarding the biosynthetic energy cost and aromatic amino acid usage of PHXPs among selected actinobacteria. The PHXPs were found to be more costly with higher usage of costly aromatic amino acids except in case of animal host associated actinobacteria. This property may be directly related to the niche specificity or niche adaptability of considered microbes.

Furthermore, our result also supports the hypothesis of Akashi and Gojobori (2002) that, the theory of energy cost is not at all fixed and may change with different metabolic pathways, translational efficiency, nutritional availability, and other environmental conditions where microbes adapt.

4.3 Comparative analysis of selected Actinobacteria

4.3.1. Comparative genomics of Amycolatopsis

4.3.1.1 Overall codon and amino acid usage analysis

Six strains of *Amycolatopsis* genus whose whole genome sequences were finished and available in public domain database at the time of this study were selected for this investigation (Supplementary file 2). Codon usage analysis (Supplementary figure 7a) of these strains clearly indicated towards the biased use of GC rich codon (color code of the heatmap has been indicated with the figure). This result also supports the effect of compositional constrains on this genus which has been previously described at section 4.1. Amino acid usage analysis on the considered strains was also explored (Supplementary figure 7b). It was observed that, Glycine, alanine, valine, leucine were the most preferred amino acids. These amino acids are of low energy cost amino acids (Akashi and Gojobori 2002). This is quite interesting. In section 4.2 we have discussed that, free living soil actinobacteria use high cost aromatic amino acids (phenylalanine, tyrosine and tryptophan) in their PHX proteins. However, while analyzing the amino acid usage of the whole proteome it is seen that, usage of aromatic amino acids is quite low. This is probably due to the very less content of costly amino acids among PMX and PLX proteins of *Amycolatopsis* which have ultimately balanced the overall energy economy of the whole proteome.

4.3.1.2 MLSA phylogenetic tree

A multilocus sequence alignment (MLSA) phylogeny was generated taking AtpI, DnaK, GyrB, FtsZ and SecA proteins (Supplementary figure 7c). *A. mediterranei* U32, *A. mediterranei* RB and *A. methanolica* 239 were in one cluster. The second cluster consisted of *A. orientalis* HCCB10007, *A. japonica* MG417CF17 and *A. lurida* NRRL 2430. This result was further validated with blast matrix analysis.

4.3.1.3 Blast matrix analysis

Supplementary figure 7d showed the blast matrix proteome analysis among selected *Amycolatopsis* strains. It was evident that, *A. mediterranei* RB and *A. mediterranei* U32 were 99.6% similar to each other. *A. japonica* and *A. orientalis* HCCB10007 shared 75.3% similarity. Following them, 70.6%

shared nature was found between *A. lurida* NRRL2430 and *A. orientalis* HCCB10007. Furthermore, *A. lurida* NRRL2430 also shared 69% identity with *A. japoniaca*. Extent of duplication within a particular proteome was also revealed from this analysis. Both *A. mediterranei* RB and *A. mediterranei* U32 were 6.9% duplicated whereas, *A. methanolica* showed 6.2% duplication within their proteome.

4.3.1.4 Average nucleotide identity (ANI) matrix

A matrix was generated based on the ANI score. Supplementary figure 7e showed a heatmap on ANI matrix. This matrix completely validates the result from Blast matrix analysis. Two strains of *A. mediterranei* (RB and U32) showed highest identity. Following them were *A. japonica* and *A. orientalis* HCCB10007 (95%).

4.3.1.5 Pan-core pool analysis

The total proteome size adding all considered proteomes constitutes pan pool whereas, the protein set shared among the selected proteomes is core pool. A total of 7839 proteins were found in the pan pool whereas the core pool was 2998 (Supplementary figure 7f).This is quite expected since with addition of proteomes, the pan pool increase and the core pool deceases. A separate phylogenetic tree was generated with the core pool (Supplementary figure 7g). Interestingly, this tree was identical to the results obtained from both MLSA and ANI matrix based phylogeny.

4.3.1.6 Cluster of ortholog (COG) analysis

Supplementary figure 7h displayed a heatmap generated on the COG classification of considered *Amycolatopsis*. J (Translation), K (Transcription), L (Replication and repair) from 'Information and storage' (IPS) category; T (Signal Transduction), M (Cell wall/ membrane/envelop biogenesis) from 'Cellular processing and signaling' (CPS) group; C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and metabolism), Q (Secondary Structure) of 'Metabolism' (Met) category were found to be the main COG categories among investigated dataset. Unknown COGs were not considered for this analysis.

4.3.2 Comparative Genomics of Bifidobacterium

4.3.2.1 Overall codon and amino acid usage analysis

Thirteen strains of *Bifidobacterium* each from different species was considered for this study (Supplementary file 2). Overall codon usage analysis revealed the effect of GC compositional constrain over all the strains (Supplementary figure 8a). G and C ending codons also predominated over the A and T rich codons. While considering the amino acid usage Glycine, alanine, proline, valine, leucine, serine and thymine were found to be preferred amino acids (Supplementary figure 8b). The overall usage of costly aromatic amino acids was low. Section 4.2 described the lower usage of costly aromatic amino acids among PHXP since *Bifidobacterium* live within human gut and obtains some major amino acids from host itself thus maintaining the energy economy.

4.3.2.2 MLSA phylogenetic tree

AtpI, DnaK, GyrB, FtsZ and SecA proteins were considered in MLSA phylogeny (Supplementary figure 8c). *Bifidobacterium angulatum* GT102, *B. adolescentis* ATCC15703 and *B. kashiwanohense* PV202 clustered together. *B. pseudolongum* PV82 and *B. animalis* subsp. Lactis DSM10140 close to each other. *B. asteroids* PRL2011 grouped with *B. coryneforme* LMG18911. *B. actinocoloniiforme* DSM22766 was fund to be close with *B. asteroids* PRL2011 and *B. coryneforme*LMG18911. Following them, *B. breve* JCM7019 and *B. longum* subsp. infantis ATCC15697 were clustered.

4.3.2.3 Blast matrix analysis

Blast matrix (Supplementary figure 8d) proteome analysis was performed on the considered strains of *Bifidibacterium*. This analysis further supported the MLSA phylogenetic clustering pattern. For instance, *B. animalis* subsp. Lactis DSM10140 and *B. pseudolongum* PV82 were found to share considerable amount of proteome size. Further, similarities among *B. asteroids* PRL2011, *B. actinocoloniiforme* DSM22766 and *B. coryneforme*LMG18911 was also evident from this analysis. MLSA clutering pattern of *B. adolescentis* ATCC15703, *B. kashiwanohense* PV202 and *B. angulatum* GT102 was also validated through blast matrix. Moreover, *B. breve* JCM7019 was found to share higher percentage of

proteins with *B. longum* subsp. infantis ATCC15697. While considering the amount of proteome duplication, *B. longum* subsp. infantis ATCC15697 showed the higher degree of proteome duplication.

4.3.2.4 Average nucleotide identity (ANI) matrix

Supplementary figure 8e displayed the ANI matrix generated on investigated *Bifidobacterium*. From this figure it is clear that, *B. longum* subsp. infantis ATCC15697 and *B. breve* JCM7019 were close to each other. *B. pseudolongum* PV82 was found to be close to *B. animalis* subsp. Lactis DSM10140. *B. coryneforme* LMG18911, *B. actinocoloniiforme* DSM22766 and *B. asteroids* PRL2011 were grouped together. The clustering pattern of *B. angulatum* GT102, *B. adolescentis* ATCC15703 and *B. kashiwanohense* PV202 was similar as found in both MLSA and blast matrix analysis.

4.3.2.5 Pan-core pool analysis

The pan and core protein pool was analyzed and showed in Supplementary figure 8f. Total 477 were found to be in core pool and 8030 were found to be in pan pool. A phylogeny was generated on core protein pool (Supplementary figure 8g). Interestingly, this tree showed the exact clustering pattern as found in MLSA and blast matrix. Thus, all the attempts made to compare these considered *Bifidobacterium* strains displayed similar results depicting the accuracy of our approach.

4.3.2.6 Cluster of Ortholog (COG) analysis

Results obtained from COG analysis of investigated *Bifidobacterium* strains has been represented as Supplementary figure 8h. It was found that J (Translation), K (Transcription) from IPS category; T (Signal Transduction), M (Cell wall/membrane/envelop biogenesis) from CPS; C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and metabolism), Q (Secondary Structure) from Met category were prevalent. Unknown COG groups were not considered.

4.3.3 Comparative genomics of Corynebacterium

4.3.3.1 Overall codon and amino acid usage

Total 38 strains of *Corynebacterium*

each from different species were considered for this study (Supplementary file 2). Supplementary figure 9a and 9b revealed the overall codon and amino acid usage of this genus. From codon usage analysis the effect of GC compositional constrain was evident. Both GC enriched codons as well as codons ending with G and C (G3 of C3 codons) was found to be preferred among all considered strains. Alanine, valine, leucine, glycine, arginine, aspartate, glutamate, serine, threonine and proline were most favored amino acids. All these amino acids are aliphatic and with low energy cost (Akashi and Gojobori 2002). Thus, the utilization of low energy cost amino acids may help *Corynebacterium* to maintain an energy economic life style.

4.3.3.2 MLSA phylogeny

Five house-keeping proteins (AtpI, DnaK, GyrB, FtsZ and SecA) were used for MLSA phylogenetic analysis. Five distinct groups were clear (Supplementary figure 9c). Clade I consisted of *Corynebacterium resistens* DSM 45100, *C. jeikeium* K411, *C. urealyticum* DSM 7109, *C. falsenii* BL8171 DSM 44353, *C. kroppenstedtii* DSM 44385, *C. lactis* RW25, *C. variabile* DSM 44702, *C. terpenotabidum* Y11. *C. ulcerans* BRAD22, *C. pseudotuberculosis* sv. Ovis P54B96, *C. diphtheriae* NCTC11397, *C. kutscheri* DSM 20755, *C. vitaeruminis* DSM 20294, *C. atypicum* R2070, *C. argentoratense* DSM 44202, *C. epidermidicanis* DSM 45586 were present in clade II. Clade III was consisted with *C. singulare* IBS B52218, *C. aurimucosum* CN1, *C. simulans* Wattiau, *C. camporealensis* DSM 44610, *C. flavescens* DSM 20296, *C. casei* LMG S19264, *C. stationis* ATCC 6872, *C. ureicelerivorans* IMMIB RIV2301, *C. imitans* DSM 44264, *C. riegelii* PUDD 83A45. *C. humireducens* NBRC 106098, *C. marinum* DSM 44953, *C. halotolerans* YIM 70093, *C.maris* DSM 45190,*C.doosanense* CAU 212,*C.testudinoris* DSM 44614, *C. uterequi*DSM 45634 were present in clade IV. Clade V was composed of *C. callunae* DSM 20147, *C. deserti* GIMN1010, *C. glutamicum* Nakagawa ATCC 13032, *C. crudilactis* JZ16 and *C. efficiens* YS314.

4.3.3.3 Blast matrix analysis

Blast matrix proteome analysis revealed the percentage of shared proteins among selected strains (Supplementary figure 9d). Deep green color represented higher similarity than

light green color. Blast matrix totally supported the clustering pattern of MLSA phylogeny. It was evident that, strains which were presented in same cluster shared more proteome similarity among each other than others. It was also revealed that, *Corynebacterium testudinoris* DSM 44614 has 40.1% (highest among considered proteomes) duplicated proteins which is far more than other investigated *Corynebacterium* strains.

4.3.3.4 ANI matrix analysis

A matrix was generated depending upon the ANI scores among all the investigated strains. Supplementary figure 9e displayed the ANI matrix. This clustering pattern of this matrix was compared with MLSA, core protein phylogeny and blast matrix and considerable similarities were found among all them.

4.3.3.5 Pan-Core pool analysis

Fig 4.16f displayed the pan core plot of considered strains. Total 561proteins were present in core pool whereas 22031 were present in pan pool. A phylogenetic tree based on core protein pool was generated (Supplementary figure 9f). Interestingly, this phylogeny perfectly matched with the results obtained from blast matrix and MLSA phylogeny further validating those findings.

To obtain an overall knowledge regarding the functionality of the protein coding genes of considered microbes COG analysis was done.

4.3.3.6 Cluster of Ortholog (COG) analysis

Supplementary figure 9h displayed a heatmap based on the COG categories among examined Corynebacterium strains. It was found that, from ISP group J (Translation), K (Transcription), L (Replication and repair) predominated. V (Intracellular trafficking and secretion), T (Signal Transduction), M (Cell wall/ membrane/envelop biogenesis), O (Post-translational modification, protein turnover, chaperone functions) from CPS category revealed considerable values. All COG categories from Met group i.e. C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), F (Nucleotide metabolism and transport), (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and metabolism) and Q (Secondary Structure) were found to be significant

among *Corynebacterium* strains selected for this study as evident from the heatmap.

4.3.4 Comparative genomics of plant host associated Frankia

4.3.4.1 Codon and Amino acid usage

Supplementary figure 10a and 10b displayed the heatmap based on the codon and amino acid usage analysis of considered *Frankia* strains. A prominent effect of GC compositional constrain was evident from the higher usage of GC enriched codons. Glycine, Serine, proline, valine, threonine, alanine were mostly used amino acids. Interestingly, in section 4.2 we have discussed that, plant host associated *Frankia* generally use high cost aromatic amino acids in their highly expressed proteins than other protein sets. However, the usage of those costly aromatic amino acids are very less among PLXPs and PMXPs which has reduced the overall usage of phenylalanine, tyrosine and tryptophan (3 costly aromatic amino acids) among investigated *Frankia* strains lowering their overall usage. This is also helping these strains to maintain an energy economic state.

4.3.4.2 MLSA phylogeny

AtpI, DnaK, GyrB, FtsZ and SecA-these five house-keeping proteins were used for phylogenetic analysis Supplementary figure 10c. Four distinct clusters were prominent. We have sub-divided cluster I into two groups Ia and Ic (resembling to the conventional clustering pattern of *Frankia* (Tisa et al. 2016). Cluster Ia was composed of *F.* alni ACN14, *Frankia* sp. AvcI1, *Frankia* sp. strain CpI1P, *Frankia* sp. strain CpI1S, *Frankia* sp. ACN1ag, *Frankia* sp. QA3. *Frankia* sp. CcI156, *Frankia* sp. strain CcI6, *Frankia* sp. CgIS1, *Frankia* sp. strain Allo2, *F. casuarinae* BR, *Frankia* sp. CgIM4, *F. casuarinae* Thr, *F. casuarinae* CcI3, *Frankia* sp. strain CeD, *Frankia* sp. strain BMG5.23, *F. casuarinae* Cg70.9, *Frankia* sp. KB5 were grouped in cluster Ic. Cluster II consisted of *Frankia* sp. BMG5 30, *F. coriariae* BMG5.1, *Candidates* Datisca Frankia glomerata Dg2. Cluster III housed *Frankia* sp. NRRL B16219, *Frankia* sp. EAN1pec, *F. discariae* BCU110501, *Frankia* sp. Cc1 17, *F. elaeagni* BMG5 12, *Frankia* sp. EI5c, *Frankia* sp. R43, *Frankia* sp. CcI49, *Frankia* sp. G2, *Frankia* sp. EUN1f. Interestingly all nod-nodulating non-N_2 fixers or ineffectively nodulating *Frankia* strains were grouped in cluster

IV.

4.3.4.3 Blast Matrix analysis

Blast matrix proteome analysis among selected *Frankia* strains has been displayed as Supplementary figure 10d. The result of blast matrix completely validated the MLSA clustering pattern. The non-infective strains showed better proteomic identity among themselves rather than other infective and N_2 fixing strains. *Frankia* sp. R43 showed highest proteome duplication (4.1%) among the *Frankia* strains considered for this analysis.

4.3.4.4 ANI matrix and phylogeny

Supplementary figure 10e showed a matrix based on the ANI score. The intensity of green colour increases with increase in ANI score. This matrix gave similar result as that of blast matrix. Hence, this is very much evident that, a distinct genomic and evolutionary barrier exists between the N_2 fixing and non-N_2 fixing *Frankia* strains.

4.3.4.5 Pan-core pool analysis

Total amount of proteins of all the strains under investigation constituted pan pool whereas, the set of protein which is common among all the compared strains make the core protein pool. Supplementary figure 10f displayed the pan-core plot of selected *Frankia* strains. Total 33156 proteins were found in pan pool and 1095 were found in core pool. A phylogenetic tree based on the core pool was generated (Supplementary figure 10g) and compared with the result obtained from MLSA and blast matrix analysis. The grouping pattern of *Frankia* was very similar in all these mentioned analyses further validating the classification pattern.

Further we did a COG analysis to enrich our knowledge.

4.3.4.6 COG analysis

Supplementary figure 10h showed a heatmap based on the COG analysis of considered *Frankia* strains. J (Translation), K (Transcription), L (Replication and repair) from IPS group; T (Signal Transduction), M (Cell wall/membrane/envelop biogenesis) from CPS category and C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), H (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and metabolism) from Met set were found to be predominant COG categories.

4.3.5 Comparative genomics among Nocardia

4.3.5.1 Overall codon and amino acid usage

Five different strains of *Nocardia*, each from different species were considered for this analysis. Supplementary figure 11a and 11b displayed the overall codon and amino acid usage of selected *Nocardia* strains. Just like other actinoabcterial genus described previously, the persistence of GC compositional constrain is very clear from the codon usage heatmap. The amino acid usage pattern was also similar to most of the other Actinobacterial genera. Alanine, glycine, valine, proline, serine, threonine, aspartate, glutamate, arginine were found to be mostly used as evident from the amino acid usage heatmap.

4.3.5.2 MLSA phylogeny analysis

The MLSA phylogenetic tree was generated taking five housekeeping proteins (AtpI, DnaK, GyrB, FtsZ and SecA) (Supplementary figure 11c). *Nocardia cyriacigeorgica* GUH2 and *Nocardia farcinica* NCTC11134 were closer to each other rather than other three strains i.e. *Nocardia nova* SH22a, *Nocardia seriolae* EM150506 and *Nocardia brasiliensis* ATCC 700358.

4.3.5.3 Blast matrix analysis

Supplementary figure 11d displayed the blast matrix analysis result of considered *Nocardia* strains. *N. cyriacigeorgica* GUH2 and *N. farcinica* NCTC11134 were found to share highest percentage of proteomic similarity rather than other strain. This also supports the MLSA phylogeny clustering pattern. *N. nova* SH22a was found to have 6.6% proteome duplication. *N. seriolae* EM150506 and *N. brasiliensis* ATCC 700358 both had 6.2% proteome duplication rate. The duplication percentage was a bit lower in *N. cyriacigeorgica* GUH2 (5.1%) and *N. farcinica* NCTC11134 (5.2%).

4.3.5.4 ANI matrix and phylogeny

The matrix based on the ANI scores among investigated *Nocardia* strains has been given in Supplementary figure 11e. The ANI matrix result was compared with blast matrix result and a similar pattern was obtained.

4.3.5.5 Pan Core protein analysis

Supplementary figure 11f showed the pan-core plot of considered *Nocardia* strains. Total 2505 proteins were in core set whereas, 16146 proteins were in pan set. A phylogenetic tree based

on the core protein set (Supplementary figure 11g) displayed similar grouping among selected *Nocardia* strains as obtained in MLSA tree, further supporting the results of MLSA and blast matrix.

4.3.5.6 COG analysis

Supplementary figure 11h displayed the major COG categories of considered *Nocardia* strains. It was revealed that, J (Translation), K (Transcription), L (Replication and repair) from IPS category; T (Signal Transduction), M (Cell wall/ membrane/envelop biogenesis), N (Cell motility) from CPS group; C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), H (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and metabolism) and Q (Secondary Structure) from Met category were prevalent among all the investigated *Nocardia* strains.

4.3.6 Comparative genomics of Rhodococcus

4.3.6.1 Codon and amino acid usage analysis

Seven *Rhodococcus* strains each from different species were selected for this analysis (Supplementary file 2). Supplementary figure 12a and 12b displayed the overall codon and amino acid usage of selected *Rhodococcus* strains. Persistence of GC compositional constrains, dominance of G or C at the third position of the codons was evident. Low cost aliphatic amino acids like alanine, valine, proline, glycine, threonine, serine, leucine, aspartic acid, glutamic acids were preferred over others maintaining an overall energy economic state of *Rhodococcus*.

4.3.6.2 MLSA phylogeny

Five housekeeping proteins (AtpI, DnaK, GyrB, FtsZ and SecA) were used for MLSA phylogenetic analysis (Supplementary figure 12c). Two distinct clusters were evident from the phylogeny. *R. aetherivorans* IcdP1, *R. pyridinivorans* SB3094, *R. equi* 103S, *R. erythropolis* BG43, *R. fascians* D188 were grouped in one cluster. *R. jostii* RHA1 and *R. opacus* B4 were separately clustered in another clade.

4.3.6.3 Blast matrix analysis

Supplementary figure 12d revealed the blast matrix proteome analysis among considered *Rhodococcus* strains. *R. jostii* RHA1 and *R. opacus* B4 shared 51.4% proteome with each other. On

other side, *R. aetherivorans* IcdP1 and *R. pyridinivorans* SB3094 shared 45.2% proteome with each other. This result further supports the clustering pattern of MLSA phylogeny.

4.3.6.4 ANI matrix analysis

The ANI matrix has been displayed in Supplementary figure 12e. The results obtained from ANI matrix completely supported blast matrix. It was observed that, this matrix gave similar clustering pattern among *Rhodococcus* strains selected for this comparative genomics study.

Further, a detailed pan-core genome analysis was done.

4.3.6.5 Pan-core pool analysis

Supplementary figure 12f displayed the pan core plot of selected *Rhodococcus* strains. Total 16510 proteins were found in pan pool and 1891 were found in core pool. Core protein based phylogenetic tree resembled enough with MLSA phylogenetic tree as well as blast matrix result further validating those results Supplementary figure 12g .

4.3.6.6 COG analysis

Supplementary figure 12h showed a heatmap based on the overall COG categories identified among investigated *Rhodococcus* strains. J (Translation), K (Transcription) from IPS category; T (Signal Transduction), M (Cell wall/membrane/envelop biogenesis) from CPS group; C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), H (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and metabolism) and Q (Secondary Structure) from Met category were found to be major COG categories among the considered strains.

4.3.7 Comparative genomics of Tropheryma

4.3.7.1 Overall codon and amino acid usage

Tropheryma is an exceptional genus among Actinobacteria since they are AT rich organisms and are associated with neurological disorders (Das et al. 2006). Another interesting fact about this genus is, only one species has been discovered till now under this genus. Hence we have considered all the 19 strains of this genus whose whole genome sequences are available in the public domain database (Supplementary file 2). Supplementary figure 13a and 13b displayed the

overall codon and amino acid usage of this genus. From the codon usage heatmap, the persistence of AT compositional constrain in clear over *Tropheryma* genus. Moreover, the dominance of T3 codons was also revealed. The amino acid usage pattern of this unusual genus was quite similar with other Actinobacterial genera. Leucine, glycine, alanine, valine, proline, serine all these low cost aliphatic amino acids were found to be mostly preferred while the usage of bulky chain aromatic amino acids (phenylalanine, tyrosine, tryptophan) was considerably less which perfectly match with the energy cost hypothesis of Gojobori and Sarkar et al.

4.3.7.2 MLSA phylogeny

Five housekeeping protein sequences (AtpI, DnaK, GyrB, FtsZ and SecA) were used for this analysis. The MLSA phylogeny divides the considered strains into two distinct groups (Supplementary figure 13c). *T. whipplei* TW08 27 and *T. whipplei* sv. 2A Twist were presented in one group while all the other 17 strains were grouped in the second cluster. This is quite interesting since all of the strains were from same species but two of them (their whole genome sequences has been completely finished while others are in permanent draft status) showed quite difference in the amino acid sequences from the rest 17 strains.

4.3.7.3 Blast matrix analysis

Supplementary figure 13d displayed the blast matrix proteome analysis. Result of this matrix totally supported the MLSA protein based phylogeny. *T. whipplei* TW08 27 and *T. whipplei* sv 2A showed 89.3% proteome similarity among each other while their percentage of identical proteins was quite low with other strains. On the contrary 90-93% proteomic similarities were observerd among the other 17 investigated strains. *T. whipplei* Neuro14 and *T. whipplei* DigADP25 showed 1.9% proteome duplication. A slightly lower (1.7%) was observed in *T. whipplei* Dig15.

4.3.7.4 ANI matrix analysis

Supplementary figure 13e shows the matrix generated on the ANI values among the 19 considered *Tropheryma* strains. The result of this analysis was a bit different from the observations we proposed from blast matrix, MLSA and pan- core protein analysis. The similarity pattern of this matrix was also different from core protein tree and MLSA tree. For instance, *T. whipplei* TW08 27 and *T. whipplei* sv

2A were not closest to each other in this case (99.6 ANI value). *T. whipplei* TW08 27 was found to be grouped with *T. whipplei* Bcu26 and *T. whipplei* Neuro1 (ANI 99.9% in both cases) whereas, *T. whipplei* sv 2A was completely separated from this group. This indicates that, TW08 27 and sv 2A strains share better proteome similarity rather than genomic identity. This may have an effect on their pathogenicity. However, more detailed study is needed in this regard.

4.3.7.5 Pan-Core protein pool analysis

The pan- core protein pool among 19 investigated *Tropheryma* strains were analysed and displayed in Supplementary figure 13f. A total of 1010 proteins were in pan pool and 716 were in core pool. A phylogenetic tree based on the core protein (Supplementary figure 13g) set again validated the results of blast matrix and MLSA phylogeny since *T. whipplei*TW08 27 and *T. whipplei* sv 2A were found to be close to each other and made a complete different cluster than others in this core protein based phylogeny too.

4.3.7.6 COG analysis

A heatmap based on the COG categories of the considered strains has been given in Supplementary figure 13h. J (Translation), K (Transcription), L (Replication and repair), B (Chromatin Structure and dynamics) from IPS category; T (Signal Transduction), M (Cell wall/ membrane/envelop biogenesis), N (Cell motility), U (Intracellular trafficking and secretion) from CPS group; C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), H (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and metabolism) and Q (Secondary Structure) from Met category were found to be major COG categories among the considered strains.

4.3.8 Comparative genomics of Streptomyces

4.3.8.1 Overall codon and amino acid usage

A total of 29 strains each from different species of *Streptomyces* was considered for this analysis. A prominent effect of GC compositional constrain was evident from the higher usage of GC enriched codons (Supplementary figure 14a). Glycine, Serine, proline, valine, threonine, alanine were mostly used amino acids

(Supplementary figure 14b). Interestingly, in section 4.2 we have discussed that, soil inhabitant *Streptomyces* generally use high cost aromatic amino acids in their highly expressed proteins than other protein sets. However, the usage of those costly aromatic amino acids are very less among PLXPs and PMXPs which has reduced the overall usage of phenylalanine, tyrosine and tryptophan (3 costly aromatic amino acids) among investigated *Streptomyces* strains lowering their overall usage. This is also helping these strains to maintain *an energy economic state.*

4.3.8.2 MLSA phylogeny

Amino acid sequences of five housekeeping proteins (AtpI, DnaK, GyrB, FtsZ and SecA) were used for generating the MLSA phylogeny of selected strains (Supplementary figure 14c). It was observed that, *Streptomyces collinus* Tu365, *Streptomyces incarnatus* NRRL8089, *Streptomyces reticuli* TUE45, *Streptomyces avermitilis* MA4680 were grouped together. Following them, *Streptomyces ambofaciens* ATCC23877, *Streptomyces parvulus* 2297, *Streptomyces lividans* TK24, *Streptomyces coelicolor* A3, *Streptomyces pactum*KLBMP 5084 were present in a cluster. *Streptomyces cyaneo griseus* noncyanogenus NMWT1, *Streptomyces leeuwenhoekii* DSM42122, *Streptomyces glaucescens* GLAO made another clade. *Streptomyces fulvissimus* DSM40593, *Streptomyces globisporus* C1027, *Streptomyces anulatus* ATCC11523, *Streptomyces griseus*subspgriseus NBRC13350, *Streptomyces flavogriseus* IAF45CD ATCC33331, *Streptomyces albus* J1074, *Streptomyces pristinaespiralis* HCCB10218 were also found to be close to each other. *Streptomyces venezuelae* Shinobu 719 in close proximity to ATCC10712, *Streptomyces vietnamensis* GIMV40001. *Streptomyces lydicus* 103 and *Streptomyces albulus* NK660 was also close to each other.

4.3.8.3 Blast matrix analysis

Blast matrix proteome analysis on selected Streptomyces strains has been presented in Supplementary figure 14d. We have compared the results from blast matrix with the MLSA phylogeny. Results were similar and thus, blast matrix further validated the MLSA phylogeny.

4.3.8.4 ANI matrix analysis

A matrix was generated depending

upon the ANI scores among all the investigated strains. Supplementary figure 14e displayed the ANI matrix. This clustering pattern of this matrix was compared with MLSA, core protein phylogeny and blast matrix and considerable similarities were found among all them.

4.3.8.5 Pan- core pool analysis

Supplementary figure 14f displayed the pan core plot of considered strains. Total 1748 proteins were present in core pool whereas 48085 were present in pan pool. A phylogenetic tree based on core protein pool was generated (Supplementary figure 14g). Interestingly, this phylogeny perfectly matched with the results obtained from blast matrix and MLSA phylogeny further validating those findings.

4.3.8.6 COG analysis

Supplementary figure 14h showed a heatmap based on the COG analysis of considered *Streptomyces* strains. J (Translation), K (Transcription), L (Replication and repair) from IPS group; D (Cell cycle control and mitosis), V (Intracellular trafficking and secretion), T (Signal Transduction), M (Cell wall/ membrane/envelop biogenesis) from CPS category and C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), H (Coenzyme metabolism), I (Lipid metabolism) from Met set were found to be predominant COG categories.

4.3.9 Comparative genomics of Propionibacterium

4.3.9.1 Overall codon and amino acid usage

Five strains of *Propionibacterium* were selected for this study (Supplementary file 2). Here we considered *Propionibacterium acnes* PA 30 2 L1, *P. acnes* HL096PA1, *P. avidum* 44067, *P. freudenreichii* shermanii JS, *P. acidipropionici* ATCC 55737, *Propionibacterium* sp. oral taxon193 F0672 and *P. propionicum* F0230a. Supplementary figure 15a and 15b displayed the overall codon and amino acid usage of selected *Propionibacterium* strains. Persistence of GC compositional constrains, dominance of G or C at the third position of the codons was evident. Low cost aliphatic amino acids like alanine, valine, proline, glycine, threonine, serine, leucine, aspartic acid, glutamic acids were preferred over others maintaining an overall energy

economic state of *Propionibacterium*.

4.3.9.2 MLSA phylogeny

Five housekeeping protein sequences (AtpI, DnaK, GyrB, FtsZ and SecA) were used for this analysis. *P. acnes* PA 30-2-L1, *P. acnes* HL096PA1, *P. avidum* 44067 and *Propionibacterium* sp. oral taxon193 F0672 were placed in one clade. Among them, the two strains of *Propionibacterium acnes* were closest (Supplementary figure 15c).

4.3.9.3 Blast matrix analysis

Blast matrix proteome analysis on selected *Propionibacterium* strains has been presented in Supplementary figure 15d. We have compared the results from blast matrix with the MLSA phylogeny. Results were similar and thus, blast matrix further validated the MLSA phylogeny.

4.3.9.4 ANI matrix analysis

A matrix was generated depending upon the ANI scores among all the investigated strains. Supplementary figure 15e displayed the ANI matrix. Color code has been given with the figure. This clustering pattern of this matrix was compared with MLSA, core protein phylogeny and blast matrix and considerable similarities were found among all them.

4.3.9.5 Pan- core pool analysis

Supplementary figure 15f displayed the pan core plot of considered strains. Total 681 proteins were present in core pool whereas 7955 were present in pan pool. A phylogenetic tree based on core protein pool was generated (Supplementary figure 15g). Interestingly, this phylogeny perfectly matched with the results obtained from blast matrix and MLSA phylogeny further validating those findings.

4.3.9.6 COG analysis

Supplementary figure 15h showed a heatmap based on the COG analysis of considered Propionibacterium strains. J (Translation), K (Transcription) from IPS group; D (Cell cycle control and mitosis), V (Intracellular trafficking and secretion), T (Signal Transduction), M (Cell wall/ membrane/envelop biogenesis), U (Intracellular trafficking and secretion), O (Post-translational modification, protein turnover, chaperone functions) from CPS category and C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), H (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and

metabolism) and Q (Secondary Structure) from Met set were found to be predominant COG categories.

4.3.10 Comparative genomics of Mycobacterium

4.3.10.1 Overall codon and amino acid usage

We selected 41 different strains of *Mycobacterium* incorporating facultative pathogens, opportunistic pathogens and non-pathogens (Supplementary file 2). Codon and amino acid usage analysis were done and have been represented as Supplementary figure 16a and 16b. GC compositional constrain was clearly evident from the heatmap. Moreover, dominance of C3 codons over G3 codons was also clear. Supplementary figure 16b displayed the overall amino acid usage among selected *Mycobacterium* strains. Glycine, alanine, valine, leucine, isoleucine, arginine, aspartate, glutamate, serine, threonine and proline were found to be preferred. As discussed previously, all these amino acids are low cost aliphatic amino acids. Thus, they may help pathogenic *Mycobacterium* in maintaining an energy economic life-style within their respective host. This finding is supported by the protein energy cost hypothesis of Akashi and Gojobori (2002). Regarding non-pathogenic strains we have discussed in section 4.2 that usage of aromatic costly amino acids were more among PHXP (Sarkar et al. 2018). However, their usage among PMXPs and PLXPs were so less that the overall use of these bulky chain amino acids were overall lower than aliphatic less energy consuming amino acids.

4.3.10.2 MLSA phylogeny

Five housekeeping proteins (AtpI, DnaK, GyrB, FtsZ and SecA) were used for constructing the MLSA phylogeny Supplementary figure 16c. A distinct clustering pattern was observed from this phylogeny. The first clade *was comprised of M. indicus pranii MTCC9506, Mycobacterium sp. MOTT36Y, M. intracellulare* MOTT64, *M. intracellulare* MOTT02, *M. chimaera* AH16, *M. avium* 104, *M. simiae* MO323, *Mycobacterium* sp. JDM601. Among these strains, *M. chimaera* AH16, *M. intracellulare* MOTT64, *M. intracellulare* MOTT02 and *M. avium* 104 were from *Mycobacterium avium* complex (MAC) (Schweickert et al. 2008). Non-pathogenic strain *M. indicus pranii MTCC9506* was found to be close to MAC. Similar result was also found by

Saha et al. (2018). Hence, this clustering pattern also supports their hypothesis that MAC has been evolved from non-pathogens and is still evolving (Saha et al.2018). *Mycobacterium simiae* MO323 and *Mycobacterium* sp. JDM601 are opportunistic pathogens (Steffani-Vallejo et al. 2018; Saha et al.2018) and have clustered with the MAC strains which are also potential pathogens.

All strains from *Mycobacterium tuberculosis* (MTB) complex were grouped in second cluster. Along with MTB strains some other potential pathogens including *Mycobacterium bovis* AF212297, *Mycobacterium africanum* GM041182, *Mycobacterium caprae* Allgaeu, *Mycobacterium canettii* CIPT140070008, *Mycobacterium microti* 12, *Mycobacterium haemophilum* DSM44634, *Mycobacterium ulcerans* Agy99, *Mycobacterium marinum* M, *Mycobacterium liflandii* 128FXT, *Mycobacterium kansasii* ATCC12478, *Mycobacterium leprae* TN were found to be clustered in this clade. Some other pathogens for instance, *Mycobacterium chelonae* CM6388 (Pinto-Gouveia et al. 2015), *Mycobacterium abscessus* (Petrini 2006), *Mycobacterium immunogenum* FLAC016 (Gupta et al. 2009), *Mycobacterium massiliense* GO 06 (Simmon et al. 2007) grouped together in the following cluster. The following clade IV was subdivided in two sub-clades. The first sub-clade was composed of non-pathogenic strains *M. vanbaalenii* PYR1, *M. gilvum* PYR GCK, *M. vaccae* 95051, *Mycobacterium* sp. MCS, *Mycobacterium* sp. KMS, *Mycobacterium* sp. JLS, *M. chubuense* NBB4, *M. phlei* CCUG21000, *M. rhodesiae* NBB3. The next sub-clade consisted of facultative pathogens which are generally found in environmental habitats but can infect immune-compromised patients. These strains were *Mycobacterium neoaurum* VKM Ac1815D (Awadh et al. 2016), *M. fortuitum* CT6, *M. smegmatis* MC2155. *Mycobacterium goodie* X7B was close to *M. smegmatis* MC2155 and *M. fortuitum* CT6 further supporting the information that, *M. goodii* are generally found in same environment as *M. smegmatis* MC2155 and *M. fortuitum* (https://en.wikipedia.org/wiki/Mycobacterium_goodii).

4.3.10.3 Blast matrix analysis

Supplementary figure 16d displayed

the blast matrix among the Mycobacterium strains considered in this analysis. The MTB strains revealed better proteomic similarity among them. Moreover, the pathogenic strains showed considerable proteomic identity among them rather than with non-pathogenic strains. Similarly, non-pathogens were more related to each other than pathogens. *M. indicus pranii MTCC9506* shared 61.3%, 84.1%, 82.3% and 80.4% proteomic similarity with *M. avium* 104, *Mycobacterium sp. MOTT36Y, M. intracellulare* MOTT64 and *M. intracellulare* MOTT02 respectively. This result completely supports the MLSA clustering pattern.

4.3.10.4 ANI matrix

A matrix was generated depending upon the ANI scores among all the investigated strains. Supplementary figure 16e displayed the ANI matrix. Color code has been indicated. This clustering pattern of this matrix was compared with MLSA, core protein phylogeny and blast matrix and considerable similarities were found among all them. *M. indicus pranii MTCC9506* was also grouped with the strains of MAC along with *M. simiae* MO323, *Mycobacterium* sp. JDM601 further validating the MLSA result.

4.3.10.5 Pan-core pool analysis

Supplementary figure 16f represents the pan-core plot of considered *Mycobacterium* strains. 28368 proteins were found to be under pan pool and 803 were in core pool. Supplementary figure 16g shows the phylogenetic tree based on core protein pool. This tree was compared with MLSA and blast matrix analysis and exactly similar trend was obtained.

4.3.10.6 COG analysis

Supplementary figure 16h showed a heatmap based on the COG analysis of considered *Propionibacterium* strains. J (Translation), K (Transcription), L (Replication and repair) from IPS group; T (Signal Transduction), M (Cell wall/membrane/envelop biogenesis) from CPS category and C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), H (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and metabolism) and Q (Secondary Structure) from Met set were found to be predominant COG categories.

The comparative genome analysis were further extended to extremophilic stone dwelling actinobacteria.

4.3.11 Comparative genomics of Blastococcus

4.3.11.1 Overall codon and amino acid usage

A total of 12 strains each from different species of *Blastococcus* were considered for this analysis (Supplementary file 2). A prominent effect of GC compositional constrain was evident from the higher usage of GC enriched codons (Supplementary figure 17a). Glycine, Serine, proline, valine, threonine, alanine were mostly used amino acids (Supplementary figure 17b). Interestingly, in section 4.2 we have discussed that, soil inhabitant *Blastococcus* generally use high cost aromatic amino acids in their highly expressed proteins than other protein sets. However, the usage of those costly aromatic amino acids are very less among PLXPs and PMXPs which has reduced the overall usage of phenylalanine, tyrosine and tryptophan (3 costly aromatic amino acids) among investigated *Blastococcus* strains lowering their overall usage. This is also helping these strains to maintain an energy economic state.

4.3.11.2 MLSA phylogeny

Amino acid sequences of five house-keeping proteins (AtpI, DnaK, GyrB, FtsZ and SecA) were used for generating the MLSA phylogeny of selected strains (Supplementary figure 17c). It was observed that, *Blastococcus* sp. DSM 44205, *Blastococcus* sp. DSM 44268, *Blastococcus* sp. AP3, *Blastococcus aggregatus* DSM 4725, *Blastococcus* sp. DSM 44270 clustered together. *Blastococcus* sp. DSM 46786 and *Blastococcus bullaregi* DSM46842 were in close proximity. *Blastococcus* sp. URHD0036, *Blastococcus endophyticus* DSM 45413 grouped with each other.

4.3.11.3 Blast matrix analysis

Blast matrix proteome analysis on selected *Blastococcus* strains has been presented in Supplementary figure 17d. We have compared the results from blast matrix with the MLSA phylogeny. Results were similar and thus, blast matrix further validated the MLSA phylogeny.

4.3.11.4 ANI matrix analysis

A matrix was generated depending upon the ANI scores among all the investigated strains. Supplementary figure 17e displayed the ANI matrix. This clustering pattern of this matrix was compared with MLSA, core protein phylogeny and blast matrix and

considerable similarities were found among all them.

4.3.11.5 Pan- core pool analysis

Supplementary figure 17f displayed the pan core plot of considered strains. Total 1877 proteins were present in core pool whereas 13269 were present in pan pool. A phylogenetic tree based on core protein pool was generated (Supplementary figure 17g). Interestingly, this phylogeny perfectly matched with the results obtained from blast matrix and MLSA phylogeny further validating those findings.

4.3.11.6 COG analysis

Supplementary figure 17h showed a heatmap based on the COG analysis of considered *Blastococcus* strains. J (Translation), K (Transcription), L (Replication and repair) from IPS group; D (Cell cycle control and mitosis), V (Intracellular trafficking and secretion), T (Signal Transduction), M (Cell wall/ membrane/envelop biogenesis) from CPS category and C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), H (Coenzyme metabolism), I (Lipid metabolism) from Met set were found to be predominant COG categories.

4.3.12 Comparative genomics of Geodermatophilus

4.3.12.1 Overall codon and amino acid usage analysis

Twenty-four strains of *Geodermatophilus* were selected for this investigation (Supplementary file 2). Codon usage analysis (Supplementary figure 17a) of these strains clearly indicated towards the biased use of GC rich codon (color code of the heatmap has been indicated with the figure). This result also supports the effect of compositional constrains on this genus which has been previously described at section 4.1. Amino acid usage analysis on the considered strains was also explored (Supplementary figure 17b). It was observed that, Glycine, alanine, valine, leucine were the most preferred amino acids. These amino acids are of low energy cost amino acids (Akashi and Gojobori 2002). This is quite interesting. In section 4.2 we have discussed that, extremophilic actinobacteria use high cost aromatic amino acids (phenylalanine, tyrosine and tryptophan) in their PHX proteins. However, while analyzing the amino acid usage of the whole proteome it is

seen that, usage of aromatic amino acids is quite low. This is probably due to the very less content of costly amino acids among PMX and PLX proteins of *Amycolatopsis* which have ultimately balanced the overall energy economy of the whole proteome.

4.3.12.2 MLSA phylogenetic tree

A multilocus sequence alignment (MLSA) phylogeny was generated taking AtpI, DnaK, GyrB, FtsZ and SecA proteins (Supplementary figure 18a). Four different clusters were found. Clade 1 comprised of *G. ruber* DSM 45317 and *G. ramli* DSM 46844. *G. obscures, G. poikilotrophi, G. siccatus, G. uthaensis, G. africanus, G. turciae, G. limestone, and G. luedemanniigrouped* in clade II.Clade III comprised of*G. aquaeductus, G. carrarae, G. amargosae, G. dictyosporus, G. saharensis, G. normandii, G. nigrescens, G. tzadiensis, G. telluris, G.* sp. DSM 44208. Clade IVwas found to be consisting of *G. soli*, *G. brasiliensis, G.* sp. Leaf369 and *G. taihuensis.*

4.3.12.3 Blast matrix analysis

Blast matrix proteome analysis on selected *Geodermatophilus* strains has been presented in Supplementary figure 18b. We have compared the results from blast matrix with the MLSA phylogeny. Results were similar and thus, blast matrix further validated the MLSA phylogeny.

4.3.12.4 ANI matrix analysis

A matrix was generated depending upon the ANI scores among all the investigated strains. Supplementary figure 18c displayed the ANI matrix. This clustering pattern of this matrix was compared with MLSA, core protein phylogeny and blast matrix and considerable similarities were found among all them.

4.3.12.5 Pan-core pool analysis

The total proteome size adding all considered proteomes constitutes pan pool whereas, the protein set shared among the selected proteomes is core pool. A total of 18091 proteins were found in the pan pool whereas the core pool was 1738 (Supplementary figure 18d).This is quite expected since with addition of proteomes, the pan pool increase and the core pool deceases. A separate phylogenetic tree was generated with the core pool (Supplementary figure 18e).This tree was compared with other phylogenomics. Interestingly, this tree was identical to the results obtained from both MLSA and ANI matrix

based phylogeny.

4.3.12.6 COG analysis

Supplementary figure 17h displayed a heatmap generated on the COG classification of considered *Geodermatophilus*. J (Translation), K (Transcription), L (Replication and repair) from 'Information and storage' (IPS) category; T (Signal Transduction), M (Cell wall/ membrane/envelop biogenesis) from 'Cellular processing and signaling' (CPS) group; C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), (Coenzyme metabolism), I (Lipid metabolism), P (Inorganic ion transport and metabolism), Q (Secondary Structure) of 'Metabolism' (Met) category were found to be the main COG categories among investigated dataset. Unknown COGs were not considered for this analysis.

4.3.13 Comparative genomics of Modestobacter

4.3.13.1 Overall codon and amino acid usage

Five strains each from different species of *Modestobacter* was considered for this analysis (Supplementary file 2). A prominent effect of GC compositional constrain was evident from the higher usage of GC enriched codons (Supplementary figure 17a). Glycine, Serine, proline, valine, threonine, alanine were mostly used amino acids (Supplementary figure 17b). Interestingly, in section 4.2 we have discussed that, soil inhabitant *Modestobacter* generally use high cost aromatic amino acids in their highly expressed proteins than other protein sets. However, the usage of those costly aromatic amino acids are very less among PLXPs and PMXPs which has reduced the overall usage of phenylalanine, tyrosine and tryptophan (3 costly aromatic amino acids) among investigated *Blastococcus* strains lowering their overall usage. This is also helping these strains to maintain an energy economic state.

4.3.13.2 MLSA phylogeny

Amino acid sequences of five housekeeping proteins (AtpI, DnaK, GyrB, FtsZ and SecA) were used for generating the MLSA phylogeny of selected strains (Supplementary figure 19a). It was observed that, *M.caceresii* KNN452b and *M. roseus* DSM45764 were placed together. *M. multiseptatus* BC501 was found to be near the previous clade. *Modestobacter* sp.

Leaf380 and *Modestobacter* sp. DSM44400 were at the lower portion of the phylogeny.

4.3.13.3 Blast matrix analysis

Blast matrix proteome analysis on selected *Modestobacter* strains has been presented in Supplementary figure 19b. We have compared the results from blast matrix with the MLSA phylogeny. Results were similar and thus, blast matrix further validated the MLSA phylogeny.

4.3.13.4 ANI matrix analysis

A matrix was generated depending upon the ANI scores among all the investigated strains. Supplementary figure 19c displayed the ANI matrix. This clustering pattern of this matrix was compared with MLSA, core protein phylogeny and blast matrix and considerable similarities were found among all them.

4.3.13.5 Pan- core pool analysis

Supplementary figure 19d displayed the pan core plot of considered strains. Total 2051 proteins were present in core pool whereas 10223 were present in pan pool. A phylogenetic tree based on core protein pool was generated (Supplementary figure 18e). Interestingly, this phylogeny perfectly matched with the results obtained from blast matrix and MLSA phylogeny further validating those findings.

4.3.13.6 COG analysis

Supplementary figure 17h showed a heatmap based on the COG analysis of considered *Modestobacter* strains. J (Translation), K (Transcription), L (Replication and repair) from IPS group; D (Cell cycle control and mitosis), V (Intracellular trafficking and secretion), T (Signal Transduction), M (Cell wall/ membrane/envelop biogenesis) from CPS category and C (Energy production and conversion), G (Carbohydrate metabolism and transport), E (Amino Acid metabolism and transport), H (Coenzyme metabolism), I (Lipid metabolism) from Met set were found to be predominant COG categories among selected *Modestobacter* strains.

4.4. Comprehensive profiling of mycobacterial secretomes

4.4.1 Enriched secretomic repertoire of Mycobacterium

Secretory proteins are of pivotal importance in cellular adhesion, communication, migration, morphogenesis, proliferation and differentiation (Ranganathan and Garg

2009; Tjalsma et al. 2004). They are also crucial in establishing the basic platform for successful interaction of pathogenic *Mycobacterium* with their respective hosts, thus, paving way for proper infection (Malen et al. 2007). In non-pathogenic strains, these extracellular protein components carry out essential functions that facilitate successful abode in environmental niche (Champion and Cox 2007; Roy et al. 2013).

Members of *Mycobacterium* were found to contain a large repository of secretory proteins for efficient cellular cross-talks execution and interaction with the host and environmental niche. The frequencies of predicted secretomes in all concerned strains have been enlisted in Table 5a. The complete sets of predicted secretomes are given in Supplementary file 7. A significant positive correlation between respective proteome size and frequency of predicted secretomes ($r = 0.84$, $P < 0.01$) among the mycobacterial members was evident from this analysis (Fig 4.8).Thus the frequency of secretomes increased with the increase in proteome size. The average secretome size of the mycobacterial pathogens (~658) was also found to be significantly lesser than that of the non-pathogens (~851) ($P < 0.01$). However, the proportion of secretomes relative to respective proteome size of organisms was nearly same for both the pathogenic (15.01%) and non-pathogenic (14.94%) members of the genus (Table 5a). Previously, Rahman et al. 2014 proposed that, non-pathogenic members of the genus *Mycobacterium* generally possess larger genomes in comparison to the pathogens. Hence, secretome size was distinctly reflecting the genome size of the mycobacterial members. Furthermore, the enriched secretomic pool among the genus *Mycobacterium* signified the strategy of efficient communication with host and environmental vicinity.

4.4.2 Amino acid usage of mycobacterial secretomes

Detailed investigation on the amino acid usage pattern of mycobacterial secretomes was carried out. Heatmap was generated based on the relative frequencies of amino acids employed by the secretory proteins (Fig 4.9a). The colour intensity gradually altered from bright red to yellow in the concerned heatmap with increase in frequency of a particular amino acid. A distinct bias towards some specific amino acids including Alanine (A),

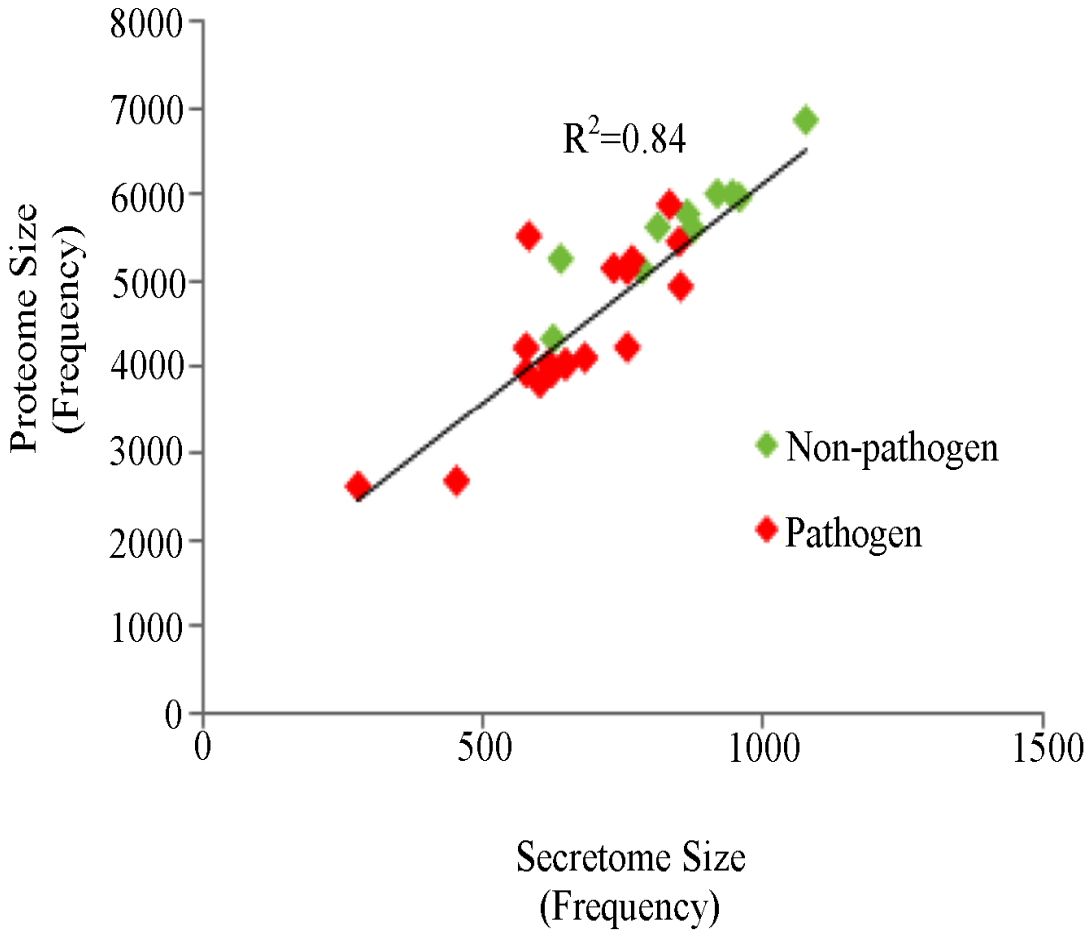


Fig 4.8 Scatter plot indicating significant positive correlation between proteome size and frequency of predicted secretomes (r = 0.84, P < 0.01) among the considered mycobacterial members.

Glycine (G), Leucine (L), Valine (V), Proline (P) and Serine (S) was evident. Secretomes of all considered *Mycobacterium* were using these aforementioned amino acids unanimously. Avoidance of Phenylalanine (F), Tyrosine (Y) and Tryptophan (W) by the secretomes was also revealed. Amino acids with bulky heavy chains for example, Arginine (R), Methionine (M) and Histidine (H) were consumed in a very low frequency by the secretory proteins. Alanine (A), Glycine (G), Proline (P) and Serine (S) were significantly more prevalent among secretory set in comparison to non-secretomic set of respective strains (Supplementary file 8). On contrary, non-secretory elements were found to be using a higher share of bulky and heavy chain amino acids like Histidine (H) and Arginine (R) in comparison to the secretory proteins ($P < 0.01$) (Fig 4.9b).

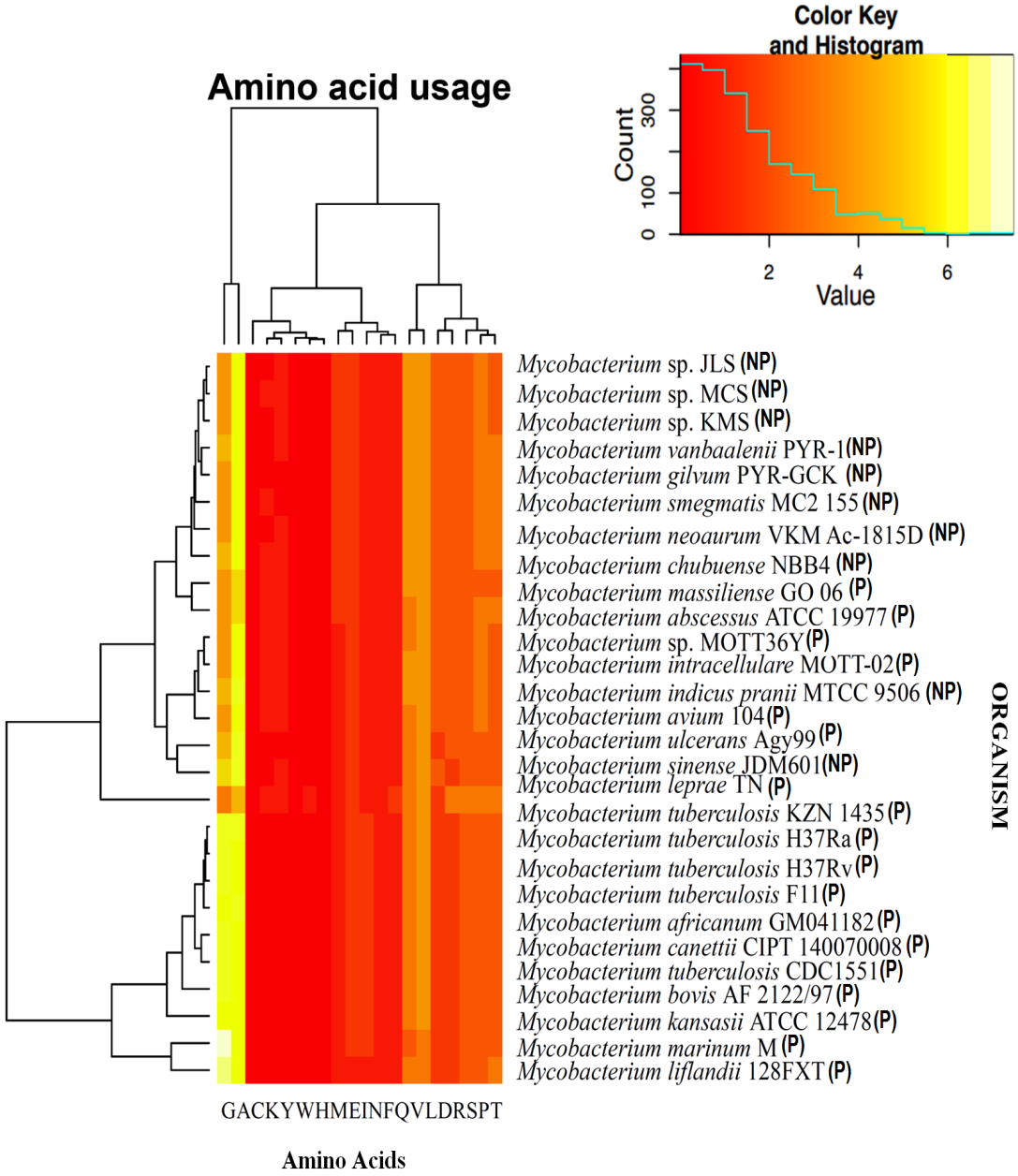


Fig 4.9a Heatmap generated based on the relative frequencies of amino acids employed by the secretory proteins.

4.4.3 Biosynthetic cost of the mycobacterial secretomes

The protein biosynthetic energy cost of secretory components was compared with that of their respective non-secretory protein set. The analysis revealed the energy economic nature of the secretomes. This may be due to the enrichment of simple and biosynthetically cheaper amino acids among secretomes along with restraint from the usage of economically costly amino acids (in terms of ATP consumption) like Histidine, Arginine and other costly aromatic amino acids. Reduced synthetic cost of secretory proteins was well justified from the

perception that, protein secretion isoften a 'one-way' mechanics due to the absence of protein importsystems (Saier 1994). Loss of extracellular proteins is irreparable and there is hardly any chance of re-cycling by the cell's chaperon and protease systems (Saier 1994). Accordingly, it is more economic and pragmatic for the microbial systems to exploit metabolically cheaper amino acids in secretomes that are permanently lost from the cell (Smith and Chapman 2010).Microbes tend to curb the biosynthetic cost of the secretory proteins for maintaining proper robustness and growth rate of the cell (Pintar and Starmer 2003).Thus, selection pressure seemed to act on the secretory components to hold back the

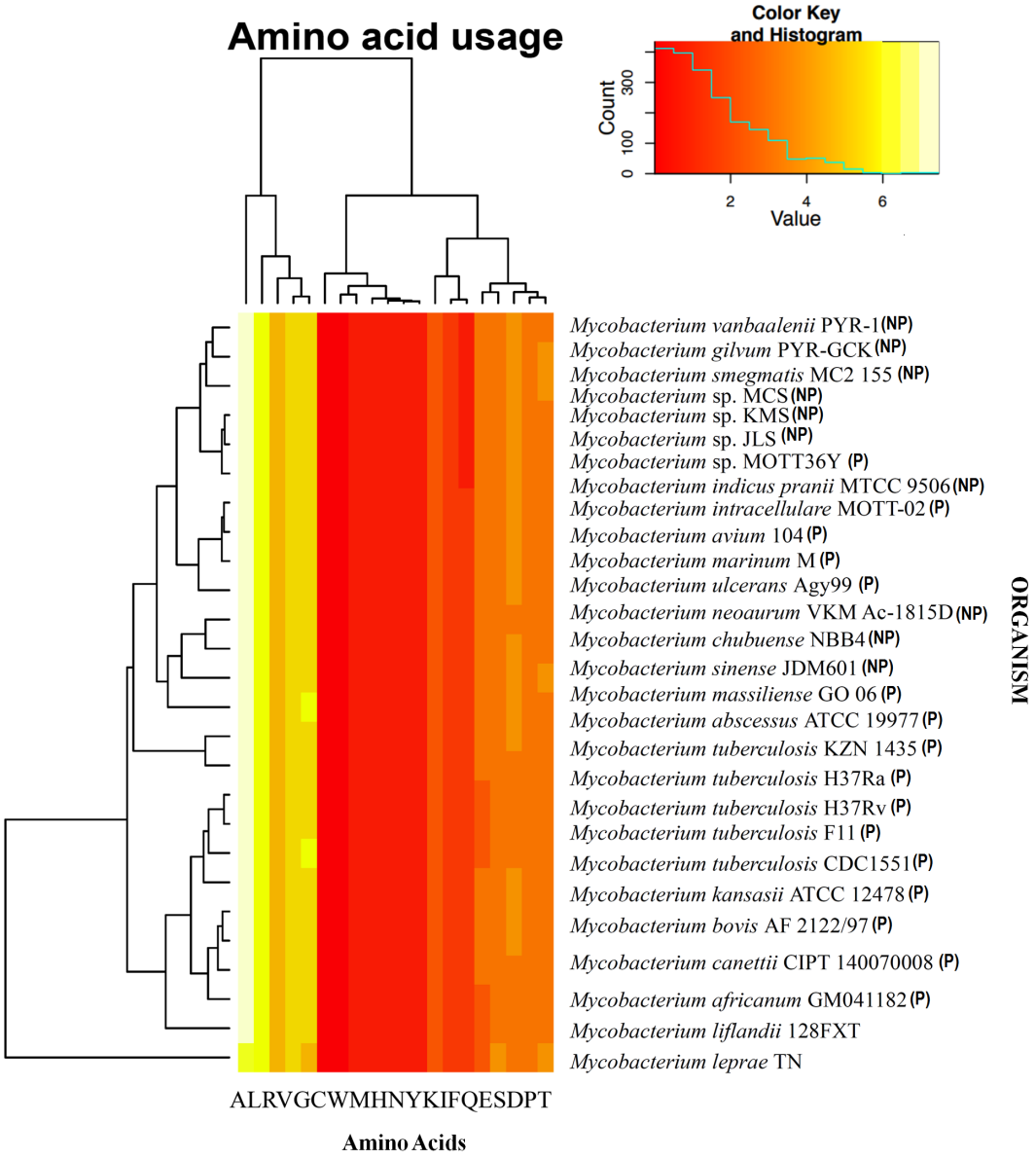


Fig 4.9b Heatmap generated based on the relative frequencies of amino acids employed by the non- secretory proteins.

cost of their production subsequently lowering the burden on the cell.

4.4.4 Functional characterization of mycobacterial secretory proteins

Detailed functional profiling of the secretomes in *Mycobacterium* was carried out by COG classification and heatmap generation based on the COG analysis (Fig 4.10a and 4.10b). A gradual change of colour shade from bright yellow to dark orange in the heatmaps indicates steady increase in frequency of a particular COG category. '*Metabolism*' (9.11%) and '*Cellular Processes and Signaling*' (14.98%) were found to be predominant among secretomes of all considered *Mycobacterium*. Hypothetical proteins and poorly characterized proteins were not considered. Further inspection revealed '*Cell wall/membrane/envelope biogenesis* [M]', '*Post-translational modification, protein turnover, chaperones* [O]' and '*Defense mechanisms* [V]' subcategories of '*Cellular Processes and Signaling*' were more frequent than others. '*Carbohydrate transport and metabolism* [G]', '*Amino acid transport and metabolism* [E]' and '*Inorganic ion transport and metabolism* [P]' sub-categories of '*Metabolism*' were also found to be prevalent. COG analysis on non-secretomic proteins was estimated and the dominance of '*Metabolism*' (38.92%) category in non secretomic set was revealed. Thus, possession of genomic architecture aiming at metabolic precision is evident from the prevalence of proteins involved in '*Metabolism*' among both secretomes and non-secretomes of investigated *Mycobacterium*. Interestingly, '*Cellular Processes and Signaling*' COG category was more predominant ($P < 0.01$) in secretome set (14.98%) rather than non secretome set (11.27%). This kind of observation seemed logical from the perspective that secretomes are more efficiently suited and devised for cell signalling and cellular cross-talks in comparison to the non-secretory components (Ranganathan and Garg 2009; Tjalsma et al. 2004).

The COG analysis of secretomes revealed that secretory proteins are mainly involved in biogenesis of cell membranes, efficient cell signaling, post-translational modifications, cellular defense and transport of carbohydrates and amino acids. This result also supports a previous report showing the effective involvement of

Organism	ISP (%)	CPS (%)	MET (%)	PC (%)
Mycobacterium tuberculosis F11 **(P)**	4.16	16.81	17.16	11.79
Mycobacterium tuberculosis H37Ra **(P)**	3.86	15.90	16.67	14.20
Mycobacterium tuberculosis H37Rv **(P)**	3.66	15.37	15.67	14.06
Mycobacterium tuberculosis KZN 1435 **(P)**	3.70	16.05	16.51	13.43
Mycobacterium tuberculosis CDC1551 **(P)**	2.42	15.42	19.06	13.34
Mycobacterium africanum GM041182 **(P)**	3.82	15.95	17.44	15.28
Mycobacterium canettii CIPT 140070008 **(P)**	3.40	16.02	16.50	13.92
Mycobacterium bovis AF 2122/97 **(P)**	3.70	15.92	16.88	15.43
Mycobacterium kansasii ATCC 12478 **(P)**	1.92	12.34	17.60	16.52
Mycobacterium leprae TN **(P)**	2.43	13.50	10.84	3.10
Mycobacterium liflandii 128FXT **(P)**	3.44	12.37	10.82	7.56
Mycobacterium marinum M **(P)**	2.00	12.44	14.08	13.26
Mycobacterium ulcerans Agy99 **(P)**	2.11	10.93	13.31	8.43
Mycobacterium avium 104 **(P)**	2.34	13.28	19.40	10.68
Mycobacterium intracellulare MOTT-02 **(P)**	3.67	14.97	20.14	11.84
Mycobacterium sp. MOTT36Y **(P)**	4.22	14.49	20.95	13.04
Mycobacterium abscessus ATCC 19977 **(P)**	2.69	12.63	16.72	7.13
Mycobacterium massiliense GO 06 **(P)**	4.36	25.82	36.36	10.54
Mycobacterium neoaurum VKM Ac-1815D **(NP)**	3.70	12.12	25.13	9.57
Mycobacterium smegmatis MC2 155 **(NP)**	3.34	12.79	26.97	8.16
Mycobacterium sinense JDM601 **(NP)**	2.40	14.72	18.24	12.96
Mycobacterium sp. JLS **(NP)**	4.04	16.74	22.17	10.05
Mycobacterium sp. KMS **(NP)**	3.48	14.78	21.20	8.80
Mycobacterium sp. MCS **(NP)**	3.69	14.27	20.79	8.98
Mycobacterium chubuense NBB4 **(NP)**	4.01	15.95	21.43	9.40
Mycobacterium gilvum PYR-GCK **(NP)**	4.11	16.67	19.75	9.47
Mycobacterium indicus pranii MTCC 9506 **(NP)**	2.19	15.67	21.94	13.95
Mycobacterium vanbaalenii PYR-1 **(NP)**	3.44	15.42	21.25	8.85

Low **High**

Fig 4.10a COG classification and heatmap generation on secretory proteins of selected *Mycobacterium* strains.

secretomes in several 'remote-control' functions that include provision of nutrients, cellular cross-talks, cell signaling and morphogenesis and are subsequently associated with the robustness of the microorganisms (Tjalsma et al. 2004).

4.4.5 Secretome based comparison of pathogenic and non-pathogenic Mycobacterium

Blast matrix (Supplementary file 9)

analysis was done to reveal the percentage of identity among the secretomic set of pathogenic and non-pathogenic *Mycobacterium* strains. The secretome set of pathogenic strains were found to share more similarity among themselves rather than their non-pathogenic counterpart. *Mycobacterium tuberculosis* complex i.e., *M. tuberculosis, M. africanum, M. bovis* and *M. canetti* were found to display high secretome-based similarity among themselves with high frequencies of shared protein families.

Organism	ISP (%)	CPS (%)	MET (%)	PC (%)
Mycobacterium tuberculosis F11 **(P)**	12.71	12.68	36.88	13.36
Mycobacterium tuberculosis H37Ra **(P)**	12.55	12.43	36.44	12.76
Mycobacterium tuberculosis H37Rv **(P)**	12.77	12.42	36.33	12.97
Mycobacterium tuberculosis KZN 1435 **(P)**	12.25	12.23	35.97	12.84
Mycobacterium tuberculosis CDC1551 **(P)**	11.90	11.87	32.90	11.76
Mycobacterium africanum GM041182 **(P)**	13.38	13.11	37.59	13.17
Mycobacterium canettii CIPT 140070008 **(P)**	12.91	12.53	36.32	12.77
Mycobacterium bovis AF 2122/97 **(P)**	12.65	12.17	36.46	12.68
Mycobacterium kansasii ATCC 12478 **(P)**	10.93	11.22	35.39	12.58
Mycobacterium leprae TN **(P)**	11.44	8.44	24.04	4.78
Mycobacterium liflandii 128FXT **(P)**	10.66	10.27	33.89	11.51
Mycobacterium marinum M **(P)**	11.92	11.52	39.98	13.29
Mycobacterium ulcerans Agy99 **(P)**	13.24	10.71	38.90	12.06
Mycobacterium avium 104 **(P)**	11.33	10.06	41.41	13.37
Mycobacterium intracellulare MOTT-02 **(P)**	12.51	10.69	44.43	14.95
Mycobacterium sp. MOTT36Y **(P)**	13.00	11.17	44.32	14.65
Mycobacterium abscessus ATCC 19977 **(P)**	14.15	9.87	39.63	12.58
Mycobacterium massiliense GO 06 **(P)**	17.02	10.06	46.72	14.00
Mycobacterium neoaurum VKM Ac-1815D **(NP)**	14.25	11.84	40.95	13.03
Mycobacterium smegmatis MC2 155 **(NP)**	12.88	9.06	43.36	13.73
Mycobacterium sinense JDM601 **(NP)**	13.04	10.83	40.83	13.25
Mycobacterium sp. JLS **(NP)**	12.60	11.32	43.22	13.76
Mycobacterium sp. KMS **(NP)**	12.21	11.17	40.19	13.04
Mycobacterium sp. MCS **(NP)**	12.42	11.24	41.47	13.37
Mycobacterium chubuense NBB4 **(NP)**	12.54	11.96	39.53	12.24
Mycobacterium gilvum PYR-GCK **(NP)**	12.29	12.01	39.94	12.60
Mycobacterium indicus pranii MTCC 9506 **(NP)**	12.61	10.77	42.45	14.53
Mycobacterium vanbaalenii PYR-1 **(NP)**	12.79	11.88	40.35	12.81

Low **High**

Fig 4.10b COG classification and heatmap generation on non-secretory proteins of selected *Mycobacterium* strains.

Members of *M. avium* complex i.e. *Mycobacterium* sp. MOTT36Y, *M. intracellulare* MOTT-02, and *M. avium* 104also showed strong similitude among them.

Interestingly, *M. intracellulare* MOTT-02, *Mycobacterium* sp. MOTT36Y and *M. avium* 104 exhibited noteworthy match with non-pathogenic *Mycobacterium indicus pranii.*This might seem logical as *M. indicus pranii* has been suggested to be the evolutionary progenitor of the pathogenic *M. avium* complex (Saini et al. 2012). High proteome-based homology between *M. intracellulare* and *M. indicus pranii*, reported earlier by Rahman and colleagues (Rahman et al. 2014), tended to support our result of considerable secretome-based similarity between the two concerned mycobacterial species. It was also evident that *M. marinum, M. ulcerans* and *M. liflandii* exhibited high degrees of resemblance with each other based on their secretory machinery. *M. leprae* was found to show feeble degrees of secretome-based similarity with the other members of the genus *Mycobacterium*. Profound gene deletion as well as considerable genome decay in *M. leprae* might be responsible for such an observation (Ore et al. 2001).

Core secretory components among pathogenic mycobacterial members were mainly found to be associated with cell elongation, ABC transporter substrate-binding, putative secreted hydrolases, peptidases and some were hypothetical proteins. Non-pathogenic members of *Mycobacterium* comprised of carboxypeptidases, hydrolases, glycosyl transferases, acyltransferases and some uncharacterized proteins in their core secretory machinery. Thus, an apparent trend was manifested that the secretory apparatus of the pathogenic members differed significantly from the secretome composition of the environmental non-pathogens of *Mycobacterium* which might accentuate their pathogenic or non-pathogenic mode of lifestyle.

4.4.6 Assessment of potentially virulent secretory proteins

Estimation of potential virulent property among the secretomes in pathogenic and non-pathogenic *Mycobacterium* revealed that the probability of secretomes being virulent is higher in pathogens (56.21%) rather than non-pathogen (49.73%) ($P < 0.01$) (Table 5b). It was observed that the in pathogens,

predicted sets of secretomes were significantly more (Mann–Whitney test, $P < 0.00001$) virulent (average proportion of potential virulent components 53.89%) than the 1000 random secretome dataset (average proportion of potential virulent components 27.70%) for all mycobacterial members. Another exciting feature was exposed depicting the fact that secretomes (for both pathogens and non-pathogens) possessed significantly higher share of potential virulent elements (56.21% for pathogens and 49.73% for non-pathogens) in comparison to the non-secretory proteins (27.01% for pathogens and 22.82% for non-pathogens) ($P < 0.01$) (Table 5b). Such a trend signifies the general tendency of mycobacterial secretomes towards being virulent.

Members of *M. tuberculosis* complex and *M. avium* complex shared substantial proportion of putative virulent proteins among their secretomes. Pathogenic microbes like *M. marinum, M. abscessus, M. ulcerans, M. liflandii* and *M.leprae* also showed high frequencies of potentially virulent secretory proteins. On the contrary, non-pathogenic mycobacterial members like *Mycobacterium* sp. KMS, *Mycobacterium* sp. JLS, *Mycobacterium* sp. MCS, *M.* sinense JDM601, *Mycobacterium neoaurum* VKM Ac-1815D and *M. smegmatis* MC2 155 were found to display low share of probable virulent secretory components. Interestingly, saprophytic *M. indicus pranii* was observed to possess a high amount of putative virulent secretory proteins which might be due to the fact that *M. indicus pranii* has been the immediate predecessor of mycobacterial members of the pathogenic *M. avium* complex (Saini et al. 2012).

In this context, it would be significant to mention that *M. indicus pranii* shares antigens with *M. tuberculosis* and *M. leprae* (Saini et al. 2012). It is often alleged to be at a state of transition towards pathogenicity due to its elevated antigenic potential (Rahman et al. 2014; Saini et al. 2012). Furthermore, it also exhibits immune modulatory activities and triggers cell mediated immune responses against *M. leprae* (Rahman et al. 2014; Saini et al. 2012).Comparative analysis with Virulence Factors Database (VFDB) revealed the presence of a large proportion of genes in *M. indicus pranii* whose orthologs have been

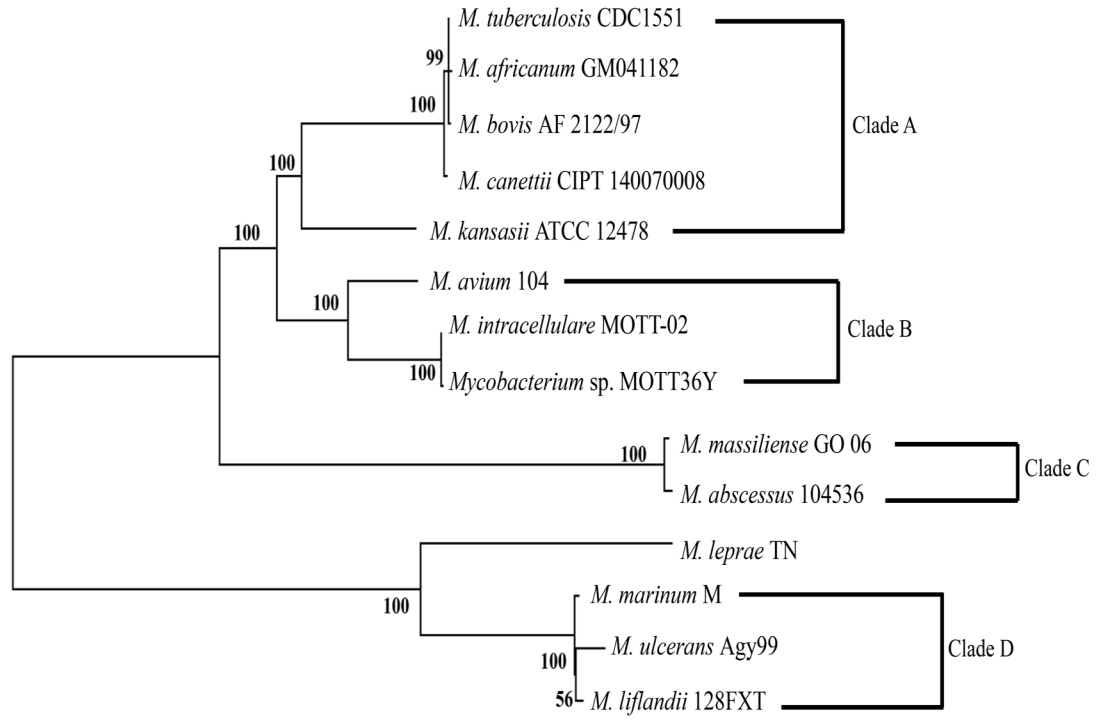
reported to be responsible for virulence in several mycobacterial pathogens (Saini et al. 2012). However, absence of certain essential elements promoting pathogenesis and infection like mce2 and mce3 operons, associated with macrophage infection (Marjanovic et al. 2010), and dev*S*/dev*R* two-component system, concerned in maintenance of dormant state of the bacterial members in depleted oxygen concentrations (Gomezet al. 2001), might be liable for the non-pathogenic nature of *M. indicus pranii*.

Mycobacterium chubuense is generally considered to be non-pathogenic (Choo et al. 2016). Surprisingly, it was found to contain an enriched pool of secretomes predicted to be potentially virulent. Such an observation tended to be in complete harmony with previous findings from Das et al. (2015) where it was reported that *M. chubuense* possesses many proteins associated with defense and virulence. Pathogenesis and infection principally depend on the ability of the infectious agent to attach, infect, survive, flourish and eventually circulate profusely inside the respective host system (Saini et al. 2012). Incompetence in any of these vital issues might be responsible for the non-pathogenic conduct of *M. chubuense*. Strikingly, non-pathogenic environmental members *M. gilvum* and *M. vanbaalenii* were also found to bear substantial proportion of predicted virulent elements among their secretomes which suggest that these environmental organisms, though non-pathogenic, might be well adapted for pathogenicity.

Since bacterial secretomes are unanimously believed to be associated with pathogenic manifestation and successful infection with host systems (Ranganathan and Garg 2009) the considerable share of probable virulent secretory components might have been responsible for the infective behavior of the mycobacterial pathogens.

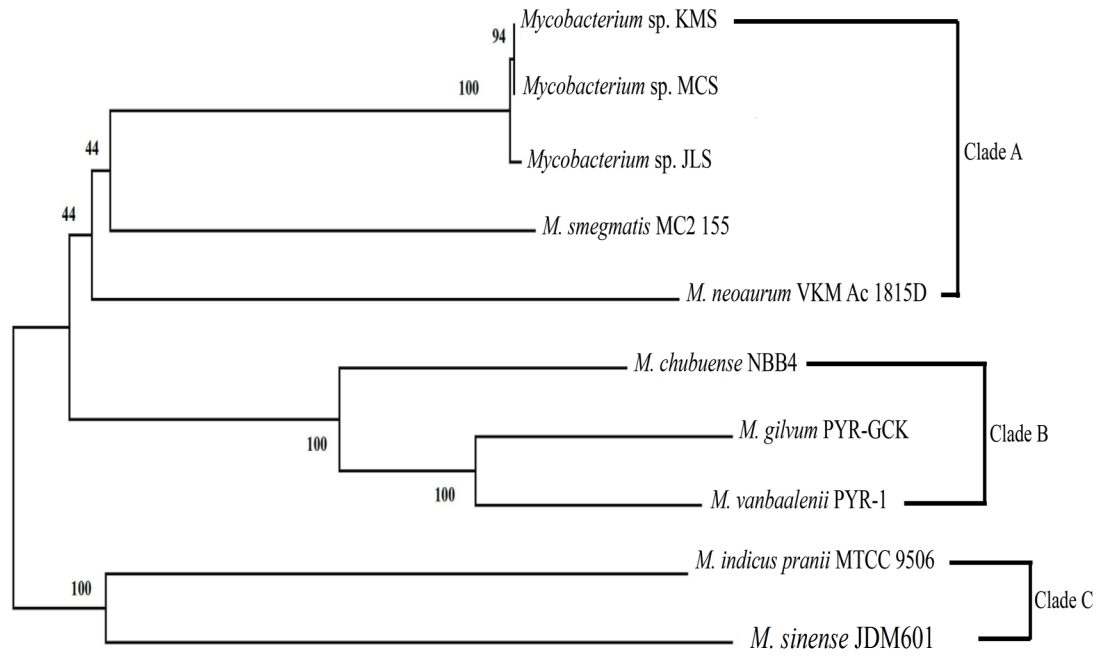
4.4.7 Evolutionary dynamics of secretory proteins

MLSA based phylogenetic trees (Fig. 4.11a, 4.11b) were generated for the both pathogenic and non-pathogenic members of *Mycobacterium* separately for performing a comparative evolutionary analysis of their secretory proteins. Phylogenetic tree based on the concatenated housekeeping proteins of the pathogenic *Mycobacterium* has been displayed in Fig. 4.11a. It was evident that there were four discrete clades (Clades A–



a.

0.02



b.

0.01

Fig 4.11 MLSA based phylogenetic trees of (a) pathogenic *Mycobacterium* (b)non-pathogenic *Mycobacterium* members considered for this study

D). Members of the pathogenic *M. tuberculosis* complex i.e., *M. africanum, M. tuberculosis, M. bovis* and *M. canetti* clustered together in Clade A. *M. kansasii* formed the outgroup in that clade. CladeB consisted of the members of *M. avium* complexi.e., M. *avium* 104, *Mycobacterium* sp. MOTT36Y and *M. intracellulare* MOTT-02. *M. avium* 104 was the outgroup member for this clade. *M.abscessus* and *M. massiliense* embodied Clade C. Clade D was comprised of *M. marinum, M. liflandii* and *M. ulcerans* where *M. leprae* was the outgroup member. This MLSA based phylogenetic tree seemed to be in complete concurrence with previous findings by Prasanna and Mehra (2013), Wang et al. (2015) and Zakham et al. (2012) and support our earlier observations pertaining to the percentage similarity analysis performed by secretome-based BLAST matrix (Supplementary file 9).

The non-pathogenic members were separated by three different clades (Clades A–C) (Fig 4.11b). *Mycobacterium* sp. KMS, *Mycobacterium* sp. MCS, *Mycobacterium* sp. JLS and *M. smegmatis*were found to group together in Clade A in which *M.neoaurum* VKM Ac-1815D was outgroup. Clade B comprised of *M. gilvum* and *M. vanbaalenii* with *M. chubuense* as the outgroup of this clade. *M. indicus pranii* and *M*. sinense JDM601 were found to represent CladeC. This result was further supported by some previous works (Prasanna and Mehra2013; Wang et al. 2015; Zakham et al. 2012).

These MLSA based phylogenetic trees for the pathogenic and non-pathogenic members of *Mycobacterium* served as the platform for evolutionary investigations. Evolutionary rate of the secretory and non-secretory protein sets of the both pathogenic and non-pathogenic mycobacterial members of each clade, was estimated with respect to the respective out group member of that particular clade. Extensive analysis displayed an appealing evolutionary pattern of the secretory proteins that tended to vary between the considered pathogenic and non-pathogenic strains. It was evident from comparative evolutionary analysis (Table 6) that the evolutionary rate of the secretory proteins in pathogenic strains of *Mycobacterium* was significantly different from the non-secretory proteins. However, such a trend was totally missing in the non-pathogens.

There was no significant difference in evolutionary rates of the secretory vs. non-secretory protein sets. Secretory elements of the pathogenic members were found to evolve faster than their respective non-secretory components and were found to be under much tranquil evolutionary constraints compared to their non-secretory counterparts.

It has already been reported that pathogenic microbes utilize cooperative secretions to adapt their environment, expand the range of their vulnerable hosts and infect multiple species (McNally et al. 2014). Rapid evolution of the extracellular secretomes has been suggested to aid bacterial adaptation (Nogueira et al. 2012). Dynamic evolutionary traits of secretory proteins of pathogenic *Mycobacterium* might award them an extra advantage to dodge the host immune responses and broaden the range of susceptible hosts to infect. Our observation of varying evolutionary features of the secretory proteins among the selected members of *Mycobacterium* might indicate the adaptive approach of the pathogens for an improved interaction and cross-talk with various range of host community. Relaxed selection forces on the secretory proteins of the pathogenic mycobacterial members might offer an opportunity to the pathogens for pertinent communication with host cellular environment and successful avoidance of host immune responses by accumulation of substitutions.

4.4.8 Contrasted evolution attributes of secretomes in Mycobacterium

Secretory proteins are composed of a two parts – a signal peptide part and a mature region (Tjalsma et al. 2004). The signal peptide part plays pivotal role protein translocation to the target cellular compartment by traversing the membrane barrier. After translocation the signal peptide part is cleaved from the respective mature part by definite signal peptidases (von Heijne 1990). It is commonly believed that the signal peptide parts perform vital functions those are unrelated to the functions of the rest of the protein (Hegde 2002). After facilitating protein translocation, the cleaved signal peptide parts act as hormones, self-antigens and neurotransmitters (Hegde 2002). Thus, signal peptides perform an array of crucial functions associated with the cellular robustness.

Depending upon the functional constraint the impact of evolutionary

forces might vary across different regions of a gene (Graur and Li 2002). To explore whether evolutionary impact varies across different regions of the secretory proteins the evolutionary rates of the signal peptide and mature parts of secretory proteins were estimated among all concerned mycobacterial members.It was evident that the signal peptide parts were evolving significantly faster than the mature parts. It suggested relaxed evolutionary constraints on signal parts (Table 6). This was consistent trend among all considered mycobacterial strains irrespective of pathogens as well as non-pathogens.

Similar observations have earlier been reported in both prokaryotic and eukaryotic systems (Buggiotti and Primmer 2006; Li et al. 2009).The rapid evolution of the functionally vibrant signal peptides appeared as a consequence of relaxed purifying selection on both the non-synonymous and synonymous sites. This might be directed towards increased translational accuracy (Li et al. 2009)improving the proficient tactics of the microbes to polish their secretion efficacy as well as communication skill with their host or environmental niche where they dwell.

4.5. Secretome profiling of extremophilic actinobacteria

4.5.1 General characterization of Geodermatophilaceae secretome set

Table 7 enlists the amount of the secretomic protein sets in each selected strains. Persistence of GC compositional constrain over the secretome encoding genes is clear from significant positive correlation between CAI, GC and GC3. GCC, CTG, GGC, GAC, GTC, GAG, CGC, CCG, CGG, CTC, CAG, GCG, GTG, ACC are some highly exploited codons among secretomic gene set. A significant ($p<0.01$) negative correlation between ENc and GC3 revealed the importance of factors other than GC and GC3 on the codon usage pattern of the investigated genes. Correlation matrix among some important parameters may prove to be helpful in identifying the factors governing the codon usage prototype of secretomic gene sets among considered strains. The correlation matrix has been represented in Supplementary file 10a. Deep green color codes for positive correlation, red codes for negative correlation and light green indicates significant effect (either positive or negative). CAI, tAI and Fop showed significant positive correlation implying the crucial participation of

both optimal and translationally optimal codons on the codon usage pattern of secretomes. The role of synonymous codon usage on secretomic genes was also exposed. Axis 1 and axis 2 were found to be the major axes in directing the codon usage outline among selected actinobacterial secretomes.

Significant positive correlation among CAI, energy cost (EC) and aromaticity among the secretomic protein (Supplementary file 10b) sets of all considered strains signified the higher usage of aromatic amino acids (phenylalanine, tyrosine and tryptophan) among secretomes (Fig 4.12). This may be due to the fact that, secretomic protein components utilize costly amino acids to maintain a proper lifestyle overcoming the adverse effect of extremophilic and nutrient scare environment. This finding also supports one of our previous findings (Sarkar et al. 2017).

4.5.2 Functional attributes of secretomes in Geodermatophilaceae

The COG analysis of secretomic set revealed the functional attributes of secretomes in signal transduction with external environment and adaptation (Fig 4.13). The major COG categories found in this case were- J (Translation, ribosomal structure and biogenesis), L (Replication, recombination and repair) from "*Information Storage and Processing*" group; V (Defense mechanisms), T (Signal transduction mechanisms), M (Cell wall/membrane/ envelope biogenesis), N (Cell motility), U (Intracellular trafficking, secretion, and vesicular transport), O (Posttranslational modification, protein turnover, chaperones) from "*Cellular Processes and Signalling*" category; C (Energy production and conversion),G (Carbohydrate transport and metabolism), H (Coenzyme transport and metabolism), P (Inorganic ion transport and metabolism) and Q (Secondary metabolites biosynthesis, transport and catabolism) from "*Metabolism*" cluster. "*Cellular Processes and Signalling*" was leading followed by "*Metabolism*" and "*Information Storage and Processing*" categories entailing the fundamental responsibility of secretomic proteins in cellular cross-talks, morphogenesis and cell signalling assisting the microbes in proper niche adaptation.

4.5.3 Secretome based comparative genomics of Geodermatophilaceae

The secretome based comparative genomics of Geodermatophilaceae

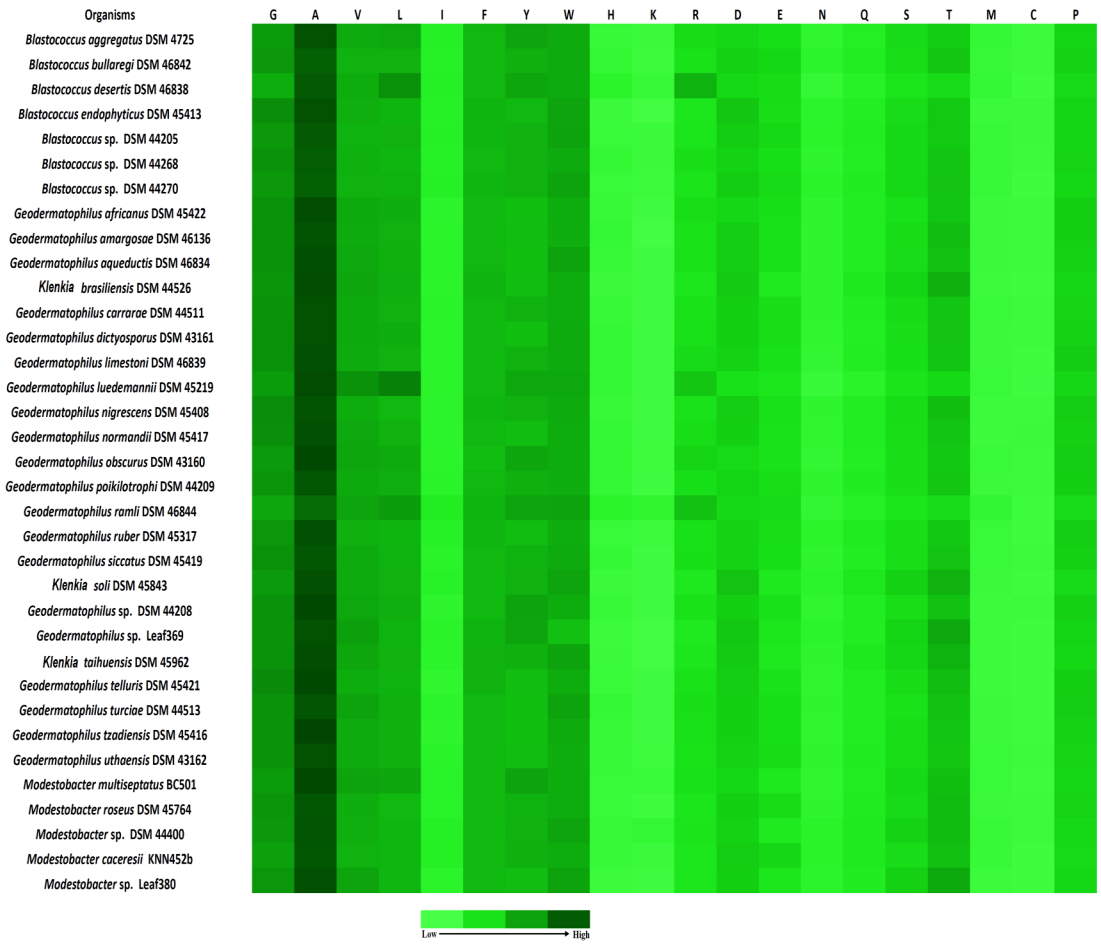


Fig 4.12 Amino acid usage heatmap of secretomes among selected Geodermatophilaceae. The color code has been indicated at the bottom of the figure.

(*Blastococcus*, *Modestobacter* and *Geodermatophilus*) was performed. Blast matrix (Fig 4.14) analysis the secretomic similarities between *B*.sp.URHD0036 and *B. endophyticus* DSM 45413 (36.1%). *B. aggregates* DSM 4725 and *B.* sp. AP3 shared 35.3% similarity among their secretomes. Around 33.3% similarity was obtained between *B.* sp. DSM44205 and *B.* sp. DSM44268. *B. desertis* showed a distinct dissimilarity in respect to other members of the genus *Blastococcus*. This may be due to the fact that *B. desertis* was isolated from desert sand whereas other strains of this genus were obtained from either soil or sea sand. While considering *Modestobacter*, *M. caceresii* KNN452b and *M. roseus* DSM 45764 were found to share higher similarity among secretomes (Fig 4.15). Among the strains of *Geodermatophilus*, *G. turciae* with *G. limstoni*; *G. uthaensis* with *G. poikilotrophi* and *G. aqueductis* with *G. carrarae* shred considerable similarities among the secretomes. Interestingly, secretomes

of *Klenkia taihuensis, K. brasiliensis* and *K. soli* (all these three strains were previously named as *Geodermatophilus* but later reclassified as *Klenkia*) displayed substantial similarities with that of *G.* sp. Leaf 369 (Fig 4.16). The results of blast matrix analysis were also supported by the pan -secretomic dendogram (Fig 4.17).

4.5.4 Evolutionary aspects

The MLSA based phylogeny (supplementary figure 17c, 18a, 19a) was used as a platform for evolutionary analysis. Evolutionary rates (dn/ds) of PHX, PLX, secretomes as well as non-secretomes of a genome were calculated with respect to the out-group of a particular clade (Table 8). This analysis revealed slower evolution of PHX genes rather than PLX genes. This is due to the fact that, organisms generally don't dare to change the essential highly expressed genes and this result also supports one of our hypothesis (Roy et al. 2015). Moreover, a detailed analysis on evolutionary pattern of secretomes and

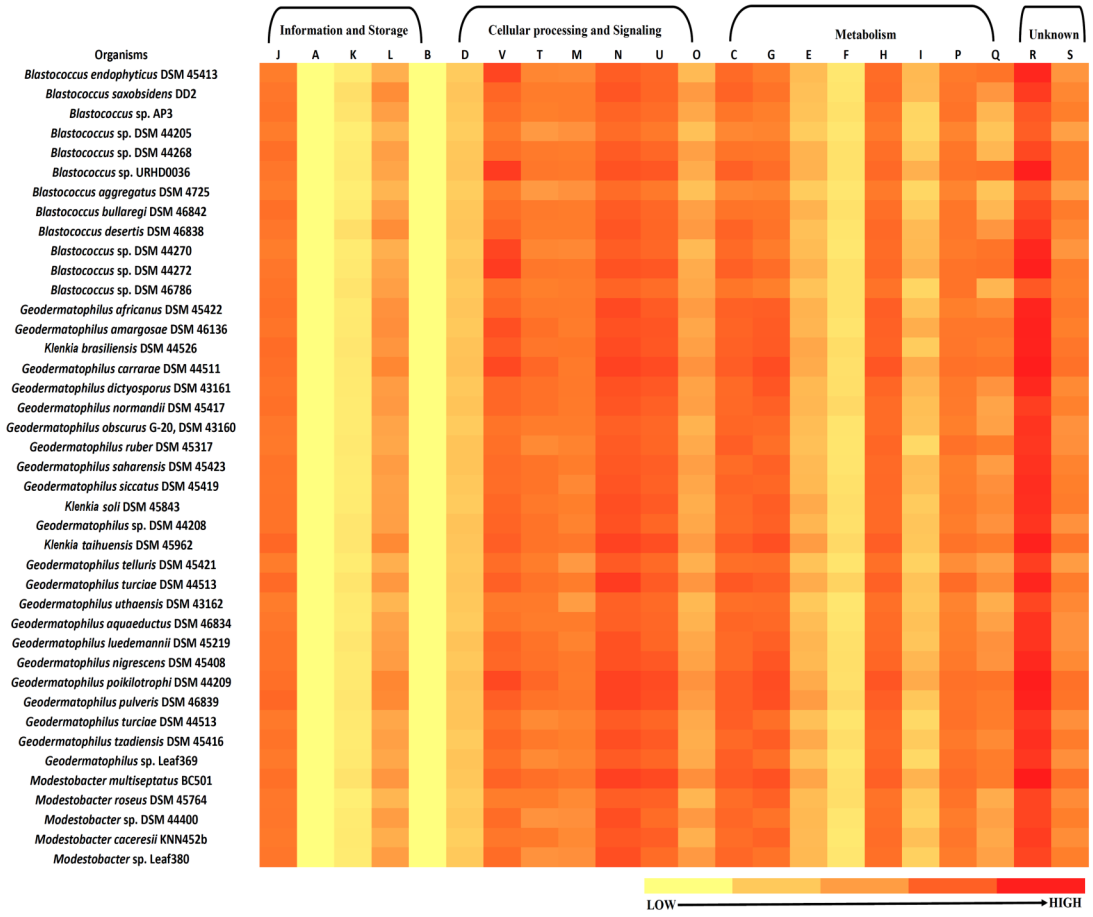


Fig 4.13 COG analysis heatmap of selected Geodermatophilaceae secretomes. The color code has been indicated at the bottom of the figure.

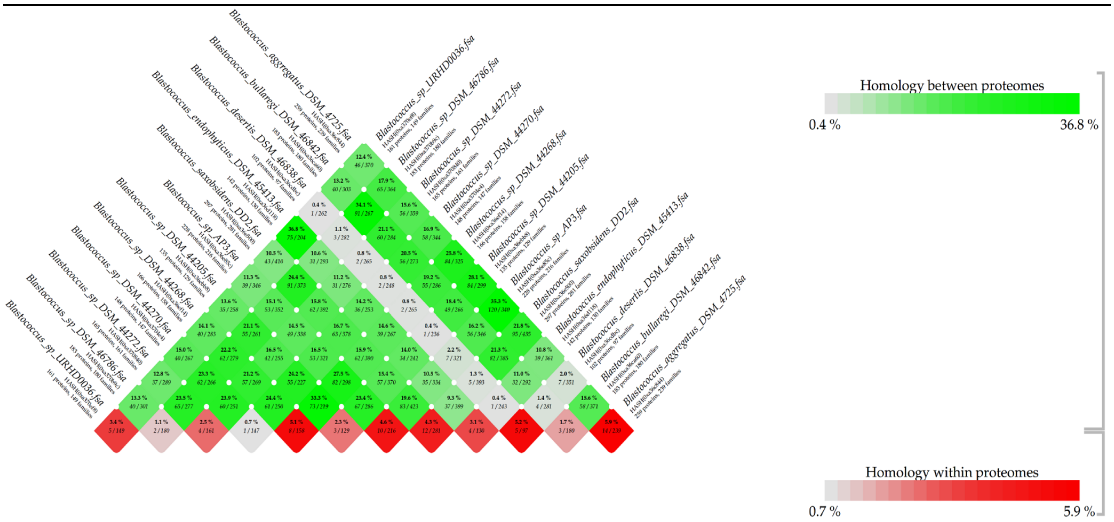


Fig 4.14 Blast matrix analysis of secretomic protein set among considered *Blastococcus* strains. Secretome set of *Blastococcus desertis* DSM46838 was found to be different from others.

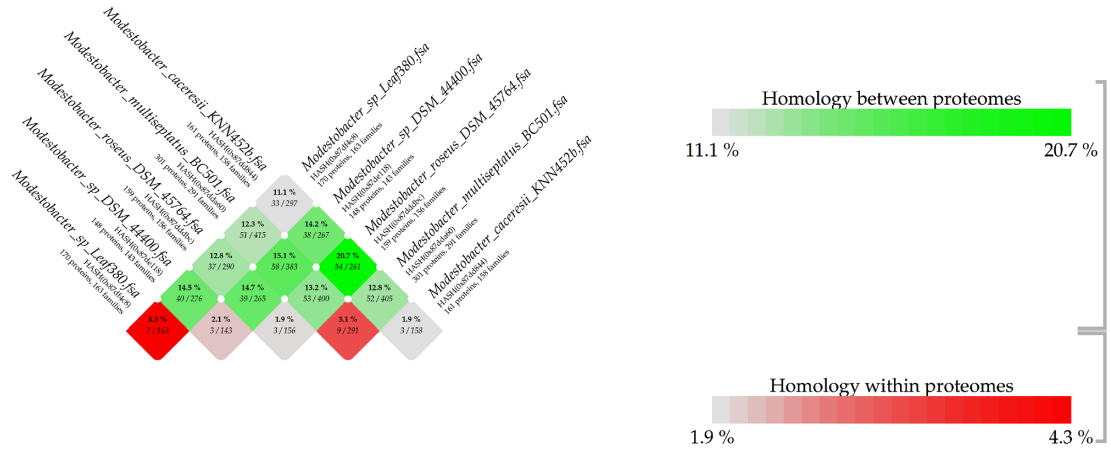


Fig 4.15 Blast matrix analysis of secretomic protein set among considered *Modestobacter* strains. Secretome set of *Modestobacter* sp. Leaf380 was found to be different from others.

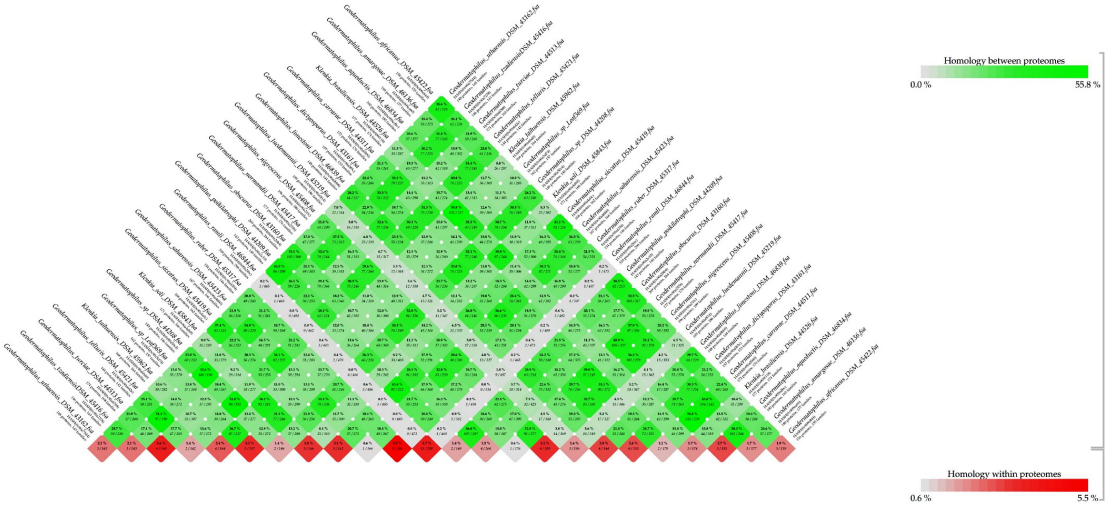
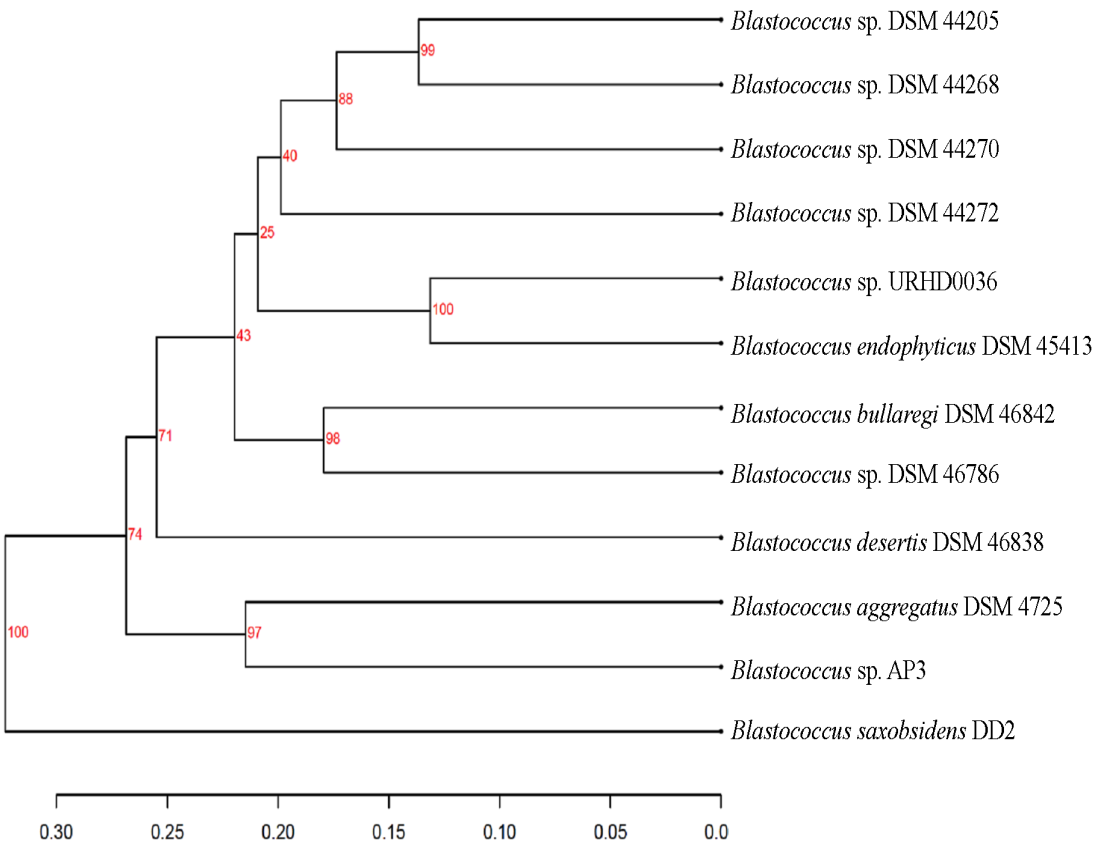
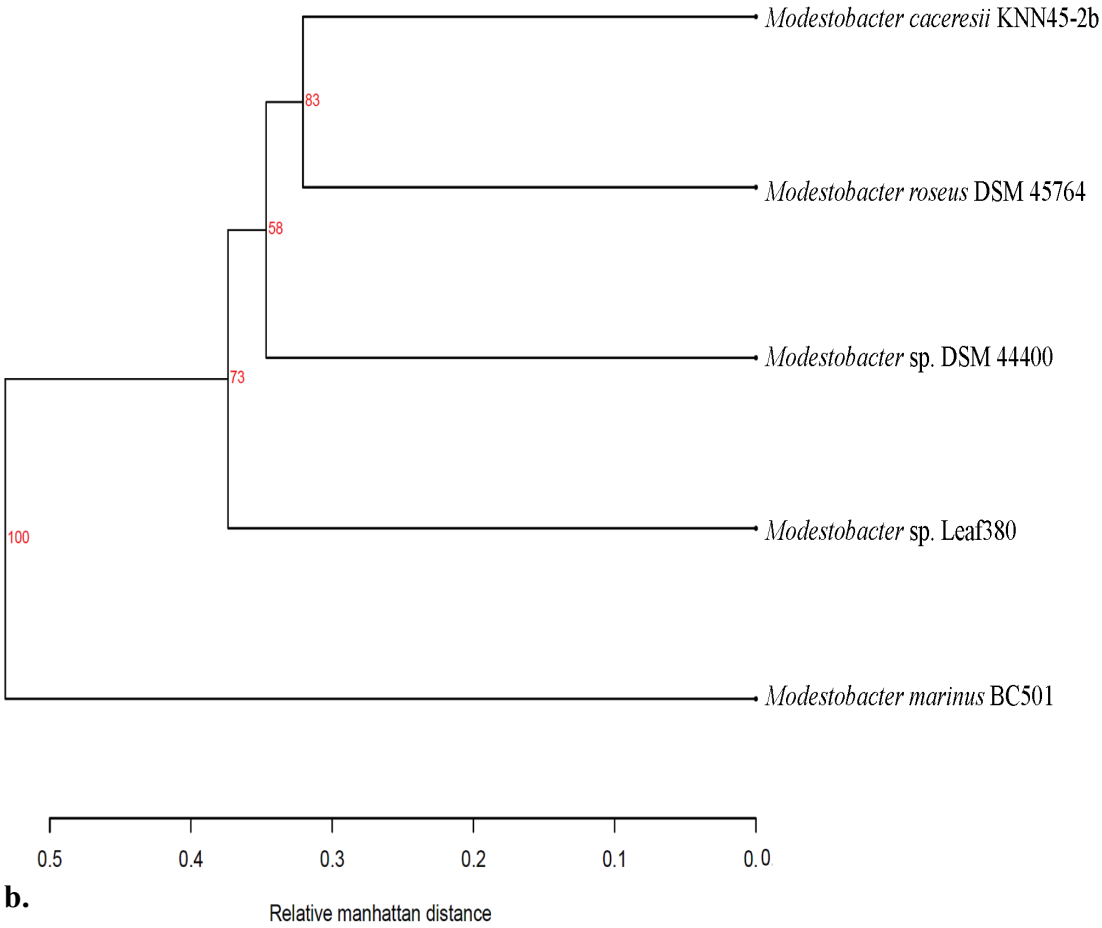


Fig 4.16 Blast matrix analysis of secretomic protein set among considered *Geodermatophilus* strains. Color code has been indicated at the right side of the figure.


Blastococcus sp. DSM 44205
Blastococcus sp. DSM 44268
Blastococcus sp. DSM 44270
Blastococcus sp. DSM 44272
Blastococcus sp. URHD0036
Blastococcus endophyticus DSM 45413
Blastococcus bullaregi DSM 46842
Blastococcus sp. DSM 46786
Blastococcus desertis DSM 46838
Blastococcus aggregatus DSM 4725
Blastococcus sp. AP3
Blastococcus saxobsidens DD2
99
88
40
25
100
43
71
98
74
100
97
0.30
0.25
0.20
0.15
0.10
0.05
0.0
Relative manhattan distance

a.


Modestobacter caceresii KNN45-2b
Modestobacter roseus DSM 45764
Modestobacter sp. DSM 44400
Modestobacter sp. Leaf380
Modestobacter marinus BC501
83
58
73
100
0.5
0.4
0.3
0.2
0.1
0.0
Relative manhattan distance

b.

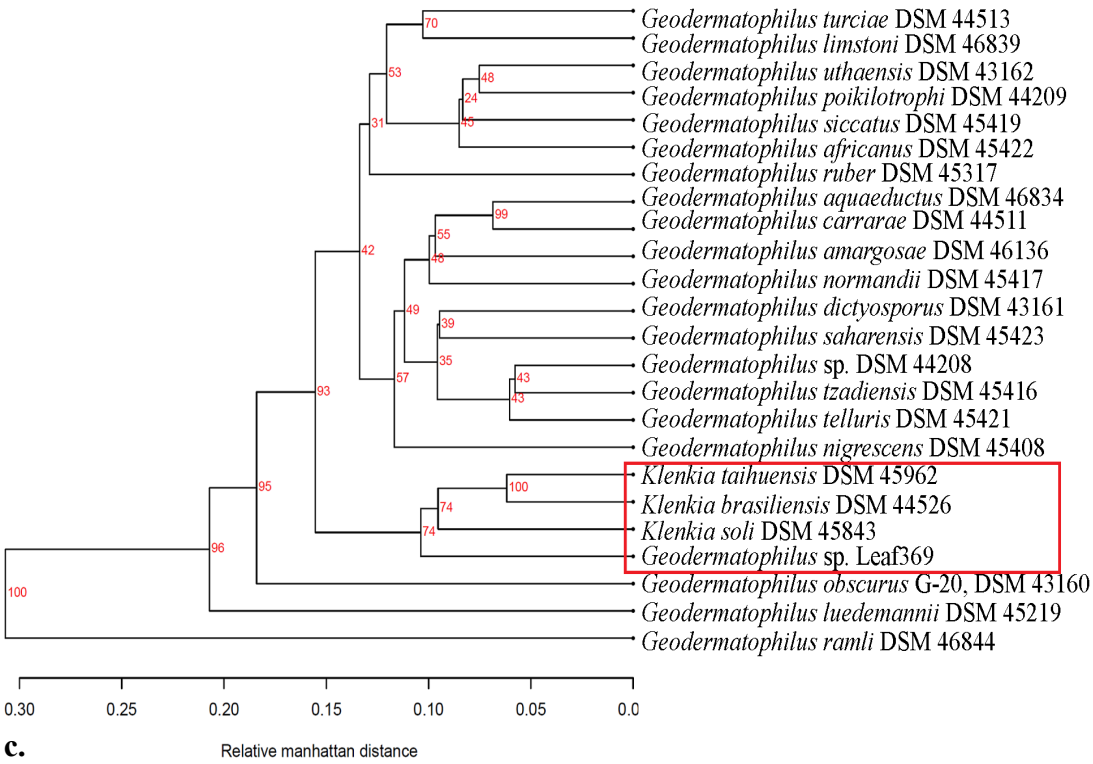


Fig 4.17 Pan-secretomic dendrogram of members of (a) *Blastococcus*, (b) *Modestobacter* and (c) *Geodermatophilus* strains selected for this study. In case of *Geodermatophilus* strains, the secretomic protein set of *Geodermatophilus* sp. Leaf 360 revealed tight clustering pattern with *Klenkia* strains than others. This same result we have also found in the comparative genomics and proteomics study discussed previously.

non-secretomes has displayed a contrasted pattern. Previously some environmental actinobacteria showed no difference between their secretome and non-secretome set however; in our case we found secretomes were evolving faster rather than the non-secretomes. This faster evolution rate of secretomes may prove to be an efficient adaptive strategy of these microbes which may further assist these bacteria in their well-organized cross talks with external environment under the stressed extremophilic condition.

4.5.5 Evolution of signal peptide parts

Assessment of evolutionary rates of signal and mature peptide parts among considered strains revealed that, signal peptide parts were evolving at a faster rate ($p<0.01$) than the mature parts (Table 8). We have previously mentioned (section 4.4.8) that, the signal peptide part is of crucial importance in protein translocation and cellular cross talks, The faster evolutionary rate on signal peptide parts indicate the persistence of relaxed purifying selection pressure on both

synonymous and non-synonymous sites. This further indicates translational precision on the signal peptide parts of investigated strains. Thus, the higher evolutionary rate of signal peptides may be related to their functions associated with cellular robustness and biological interactions. This result also supports a previous finding which states that, depending upon the functional constraint the impact of evolutionary forces might vary across different regions of a gene (Graur and Li 2002).

4.6 Biology and evolution of frankial CAZymes

4.6.1. Plant polysaccharide degrading CAZymes are predominant

The total number of CAZymes found in each *Frankia* strains has been listed in Table 9. The COG analysis of CAZymes exposed their contribution in "Metabolism" category particularly in carbohydrate metabolism (G) followed by "Cellular Processing and Signalling" (Fig 4.18, Supplementary File 11). Moreover, GH and GT types were found to be widespread over other CAZyme categories perhaps due to their crucial responsibility in plant polysaccharide degradation (Andrade et al. 2017). The plant cell wall is mainly composed of cellulose, hemicellulose, xylan, mannan, pectin, cutin, chitin, glucan, startch and other complex polysaccharides. Among them, cellulose is one of the chief plant cell wall element build up by linear chain of several β-(1-4) linked D-glucose units and can be degraded by GH-encoded cellulase enzymes. Cellulases enzyme are further categorized into endo-β-1,4-D-glucanases, cellobiohydrolases (exocellulases), and β-1,4-glucosidases (Zerillo et al. 2013) all of which are encoded by different GH member proteins. Likewise, endoglucanases and β-glucosidases are also concerned with cellulose and xyloglucan degradation. Xyloglucan is a hemicellulose present in the primary cell wall of plants. They remain associated with the cellulose microfibrils providing structural integrity to the cell wall. GH10 and GH11 are alleged to encode xyloglucan degrading enzymes. Furthermore, endo -β-1,4-xylanase and β-1,4-D-xylosidase, two well known xylan (abundant in secondary cell wall of woody plants) degraders are also encoded by GH cazymes (Zerillo et al. 2013). GH2, GH5 and GH25 have been reported to code potent mannan degraders for example β-mannosidase

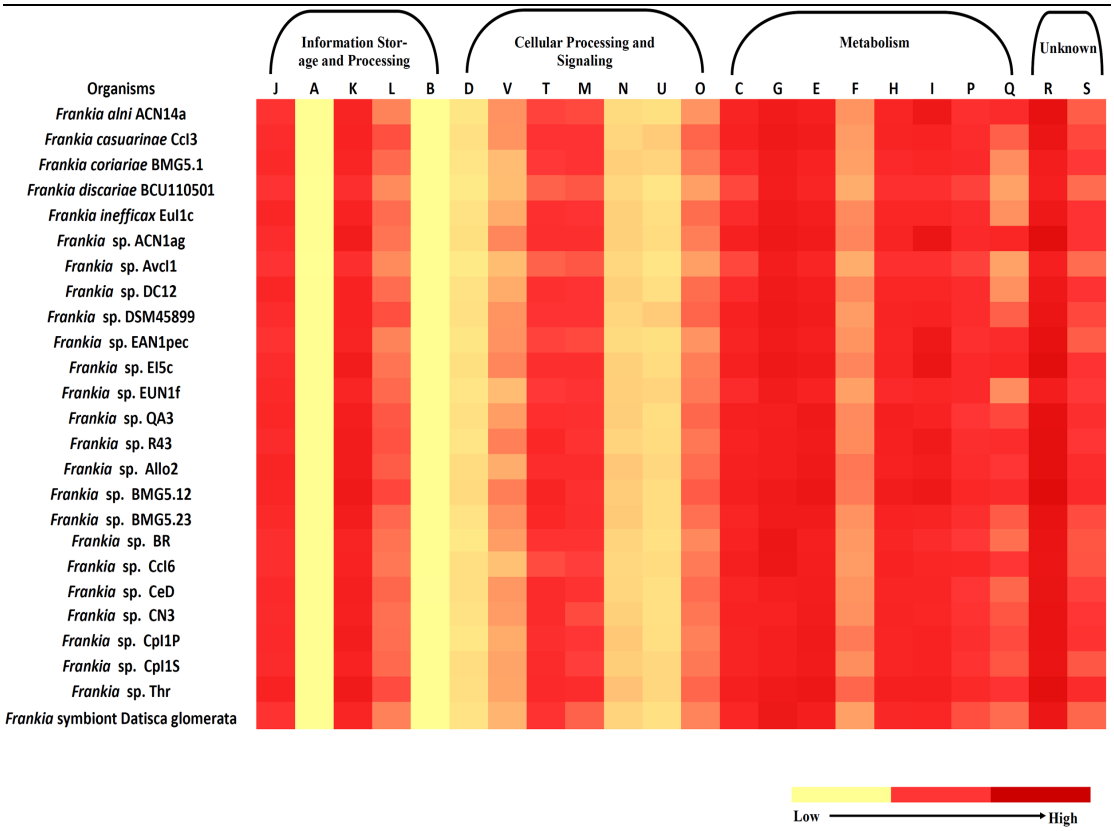


Fig 4.18: COG analysis of investigated frankial CAZyme set. Color code has been indicated at the bottom of the figure.

and β-mannase. Likewise, several other enzymes for example pectin lyase, rhamno-galacturonases, exo- and endo-polygalacturonases (pectin degrader), cutinase (degrade cutin), chitinase (involved in chitin breakdown), β-1,3-glucanosyltransglycosylases and β-1,3-glucanases (associated with the cross-linkage and alteration of linear β-1,3-glucans) are also encrypted by different GH family. Hence, the prevalence of GH among *Frankia* may help in plant polysaccharide degradation. The second most abundant CAZyme family in *Frankia* was GT. This result substantiates some previous works on soil bacteria where GH numbers were found to exceed GT numbers, whereas aquatic microbes had GT more than GH (Andrade et al. 2017). GT families are generally associated with carbohydrate biosynthesis. Both UDP-GlcNAc: peptide β-N-acetylglucosaminyl transferase and UDP-Glc:peptide N-β-glucosyltransferase are grouped as GTs and are concerned with bacterial cell membrane formation. Moreover, GTs are also associated with cell signalling, energy metabolism and stress response along with the cellulose biosynthesis, production of chitin, hyaluronic acid and sucrose (Andrade et al. 2017). Thus, frankial GHs and GTs represent

an enormous bio-resource of vital enzymes with both ecological and industrial importance. These aforementioned findings also suggest that carbohydrate-breaking enzymes in *Frankia* may aid obtaining nutrition from the plant hosts.

We tried to thoroughly investigate these highly important proteins among plant host associated *Frankia*.

4.6.2 Codon and amino acid usage analysis

A comparative account between the codon usage indices of CAZymes in each considered strains with their respective genomes revealed several crucial parameters governing codon usage pattern of this special gene set (Supplementary File 12). Perseverance of the GC compositional constrain was obvious based on the significant positive correlation ($p<0.001$) among GC, GC3 and CAI. A significant negative correlation ($p<0.001$) among GC3 and Nc exposed the effect of some parameters on CAZymes other than compositional constrain (Roy et al. 2015). To unveil those other parameters, GC, GC3 and CAI were correlated with Fop. A significant positive correlation ($p<0.001$) among these four factors indicates that Fop is another crucial factor governing codon usage model. Furthermore, Axis 1 and Axis 2 of RSCU were found to be the major axes having an effect on CAZyme codon usage (Supplementary File 12). The PHX set for each strain were thoroughly studied and a substantial amount (55-63%) of CAZymes were found to be potentially highly expressed signifying their essentiality in the frankial lifestyle. This result was further validated by analysis of the microarray data of *F. alni* ACN14a (Supplementary File 13). Analysis of the amino acid usage pattern for CAZymes revealed a significant positive correlation ($p<0.001$) among EC, CAI, and Aromaticity indicating their higher usage of aromatic amino acids (Supplementary File 14). This result supports a previous study (Sarkar et al. 2018) where the CAZyme-enriched highly expressed protein set for *Frankia* were found to utilize more biosynthetic energy due to enhanced exploitation of costly aromatic amino acids. A heatmap was generated on the basis of aromatic amino acid utilization. The heatmap clearly exposed the enrichment of aromatic amino acid by frankial CAZymes (Fig 4.19). These aromatic amino acids are

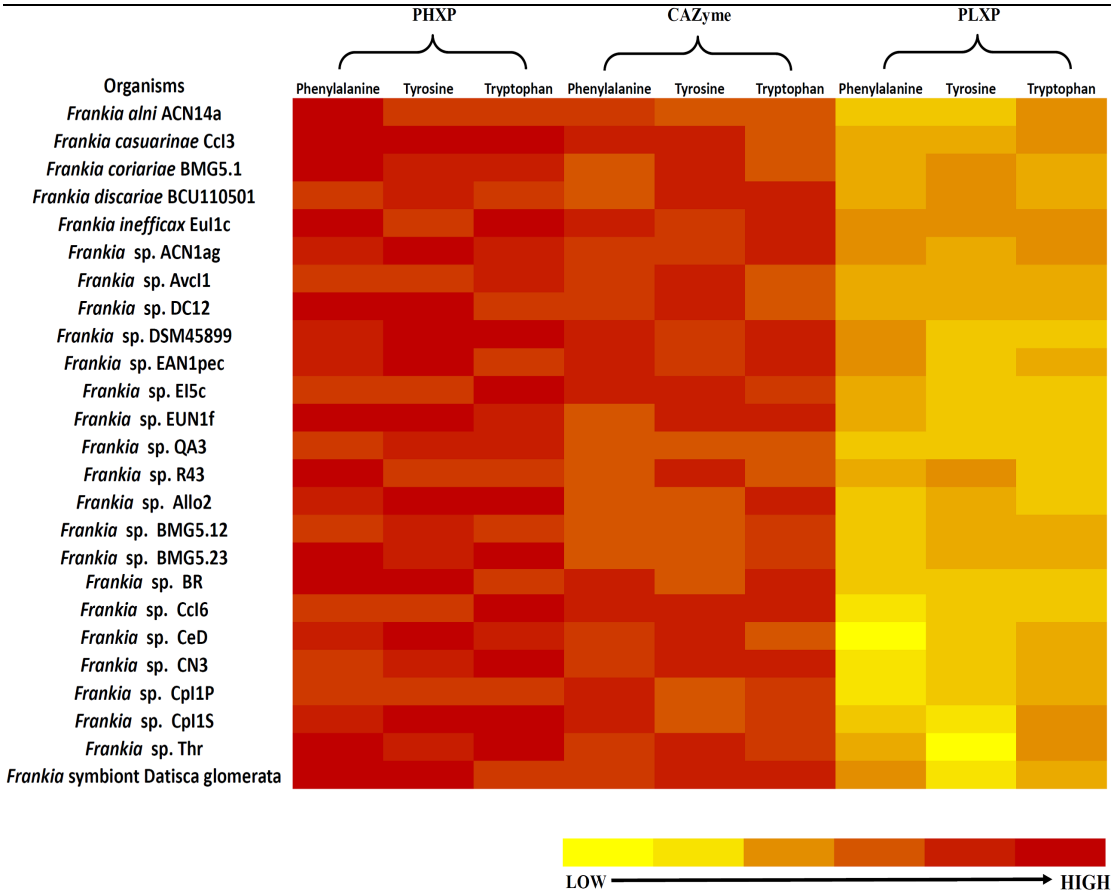


Fig4.19: Aromatic amino acid enrichment among CAZyme set of selected *Frankia*. Color code has been indicated at the bottom of the figure.

mainly present at the active sites of CAZymes and contribute to their strong binding capacity to carbohydrates moieties (Sarkar et al. 2018).

4.6.3 Phylogenomics and comparative CAZyme analysis

Previously, *Frankia* was grouped into four clusters (Ghodhbane-Gtari et al. 2010). All the four clusters were completely recovered from the MLSA phylogeny supporting the ANI matrix (Fig 4.20, 4.21). In all cases, *Frankia* strains were found to group in 4 different clusters according to their host plant range. Since CAZymes are also crucial in plant-microbe interactions, we carried out a comparative CAZyme profiling for *Frankia*. Interestingly, the blast matrix (Supplementary file 15) and core-CAZyme dendogram (Fig 4.22) based on the CAZY set of *Frankia* recovered all four *Frankia* clusters indicating an imperative role of CAZYmes in Frankial lifestyle. To further investigate whether any distinct evolutionary aspect persists over CAZymes we accomplished a

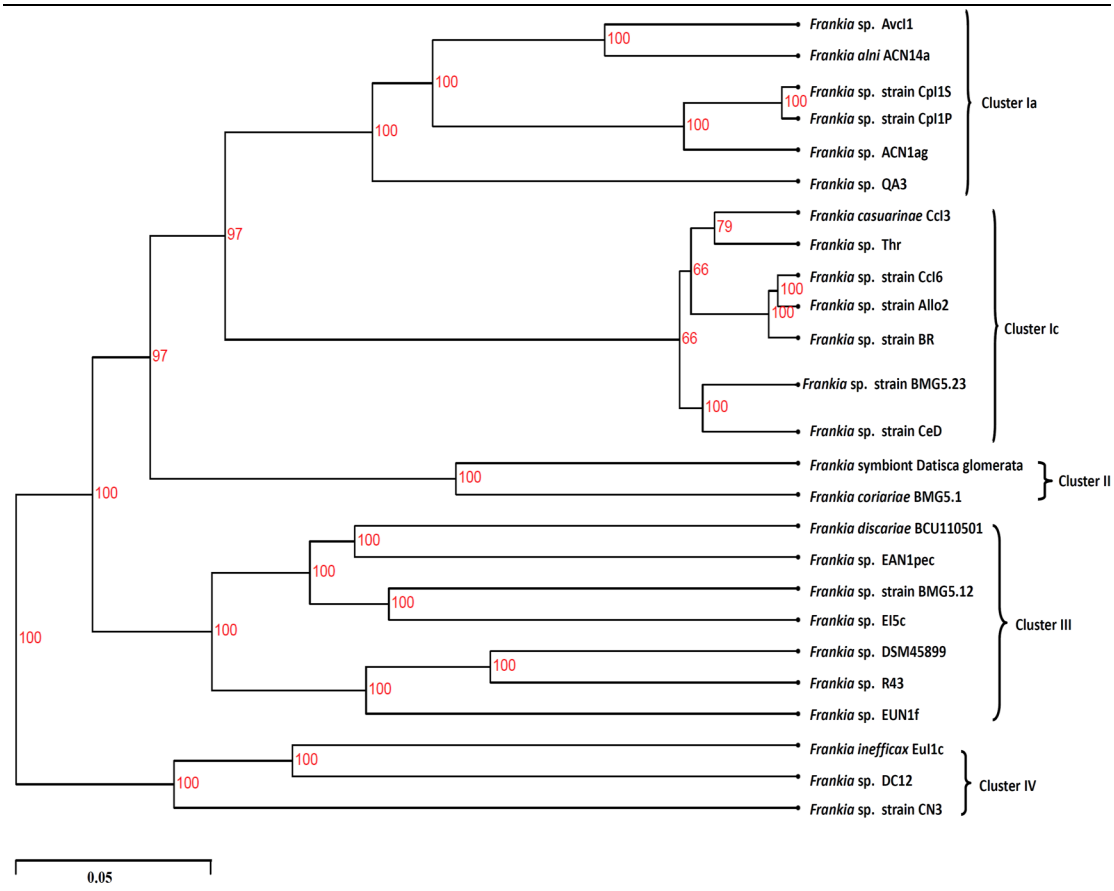


Fig4.20: MLSA phylogeny of selected *Frankia* strains using NJ method and 1000 bootstrap value.

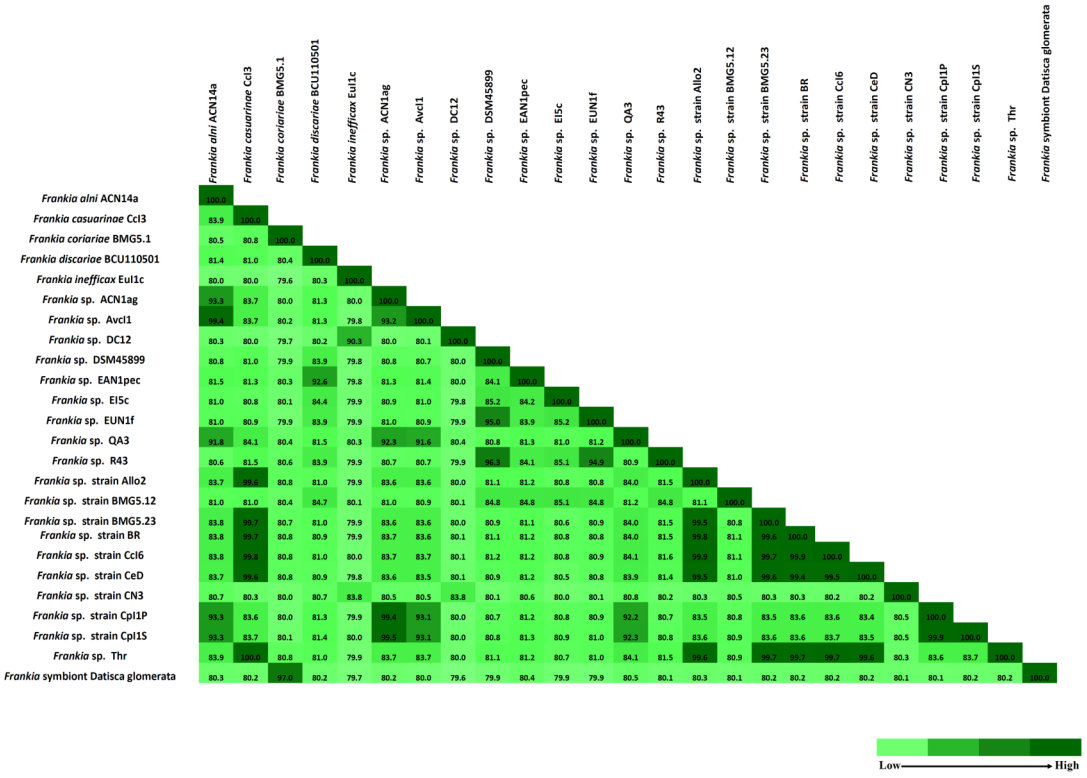


Fig 4.21: Average nucleotide identity (ANI) score matrix for considered *Frankia* strains.

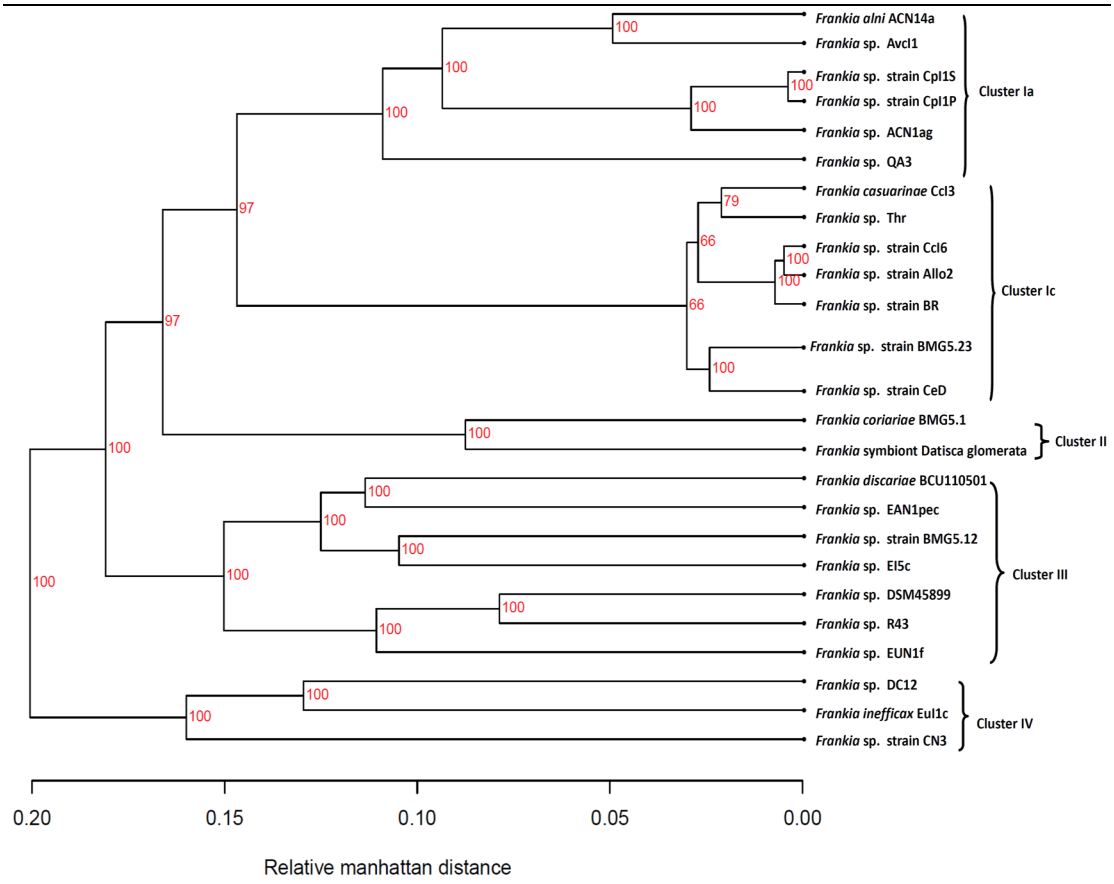


Fig 4.22: Core- cazyme based dendrogram among selected *Frankia* strains.

comprehensive evolutionary analysis.

4.6.4 Evolutionary analysis

A comparative evolutionary analysis among PHX, PLX and CAZymes for the selected *Frankia* strains was performed. Results clearly exposed that CAZyme genes were less evolving than the PLX set ($p<0.001$). The PHX genes contained a considerable amount of the CAZyme genes in all of the strains suggesting persistence of purifying selection over them (Roy et al. 2015).

However, comparison of the evolutionary rates of the CAZyme and PHX obtained appealing results (Table 10). The dN/dS values for CAZymes were found to be lower (at $p<0.001$ level) than those values for the other PHX genes among effectively nodulating *Frankia* Cluster I, II and III. In contrast, no statistical difference was observed between the PHX and CAZyme sets for non-nodulating or ineffectively nodulating *Frankia* cluster IV. The strong purifying selection on CAZymes for plant-root-colonizing strains may be due to their essentiality in plant host association as

well as effective nodulation. Since the Cluster IV *Frankia* are incompetent of nodulating plant root or nodulate ineffectively (Tisa et al. 2016), their CAZyme set showed no difference in evolutionary rate when compared to the other PHX genes. Furthermore, carbohydrates are not only vital in nodulation process. They play a critical role in N_2 fixation and assimilation process too (Gibson 1966). Thus, the impact of CAZyme set is more on nodulating, nitrogen fixing and N_2 assimilating *Frankia* strains rather than ineffectively nodulating or non-nodulating strains.

4.7 Domain based phylogenetic tree generation

4.7.1 Pan- and Core- domain set

A total of 4776 domain families were found to be accommodated in the pan domain (the total domains present in all of the organisms) category, while 342 were recognized as the core domain family (shared domains amongst all of the strains). Interconnected biological network of these domains were further studied and a well-built biological interface with PPI enrichment P value 1.0e-16 was obtained (false discovery rate < 1.00 was considered only) (Fig. 4.23). This result suggested a more significant interaction than expected (P value = 0). These domains were found to be associated in several biosynthetic and metabolic processes along with ribosome biogenesis, DNA repair, DNA recombination, translation and gene expression (Supplementary File 16).

4.7.2 Topology of the Domain based tree

The domain-based phylogenetic tree has been represented as Fig 4.24. Genera with more than one genome (*Frankia, Bifidobacterium, Corynebacterium, Streptomyces, Arthrobacter, Mycobacterium, Rhodococcus, Rothia, Nocardiopsis*) displayed a steady clustering (Fig 4.44). We divided the whole tree in six different clades (C-I to C-VI) for the ease of discussion. The major exceptions we got in our domain based tree (when compared to the MLSA phylogeny (Sen et al. 2014) have been indicated by asterisk (*). Two thermal actinobacteria (*Acidimicrobium ferroxidens* and *Acidothermus cellulolyticus)* were grouped together in (C-I) (Fig 4.24). Further analysis revealed 1070 common domains between these two strains (Supplementary File 17a). Biological network analysis of those domains

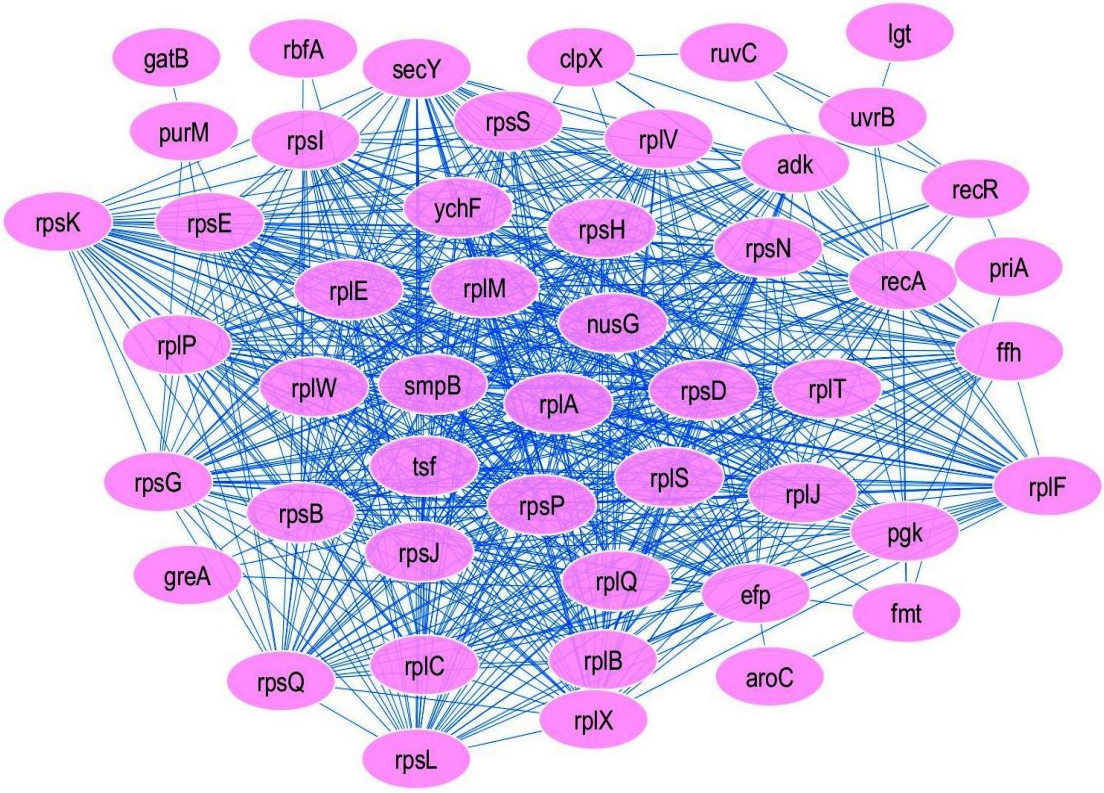


Fig 4.23 Biological interaction among selected core domains

revealed a PPI enrichment p-value of< 1.0e-[16], indicating that, those domains were biologically connected and their interaction was significantly higher than the expected value (Supplementary File 17b). Cluster II (C -II) was diverse and divided into two distinct sub-groups (C-IIa and C-IIb). In C-IIa, *Mobiluncus curtisii* and *Arcanobacterium haemolyticum* of Acidimicrobiales grouped together in one branch, *Tropheryma whipplei* (member of Micrococcales) grouped together with strains of *Rothia* in the next branch of C-IIa. Members of *Propionibacterium* also grouped with C-IIa. The *Bifidobacterium* strains were clustered jointly in C-IIb. *Gardnerella vaginalis*, clustered along with *Bifidobacterium* in our tree. This pattern was also found in case of MLSA tree (Sen et al. 2014). The subsequent cluster (C-III) was also separated into C-IIIa and C-IIIb parts. C-IIIa was comprised mainly of Micrococcales group. However, an extremophilic soil actinobacteria *Kineococcus radiotolerans*, was placed near aforementioned Micrococcales group in C-IIIa in our tree. Though, *K. radiotolerans* is able to tolerate elevated amount of radiation, they are fundamentally soil bacteria. A total of 1294 domains were found to be shared among *K. radiotolerans, J. denitrificans* and *S. keddieii*

(Supplementary File 18a). These domains were mainly associated with metabolism and cellular signaling. Biological network analysis with these common domains displayed significant more interaction than expected with a PPI enrichment p-value of<$1.0e^{-16}$ (Supplementary File 18b). Thus, the strange clustering pattern of *K. radiotolerans* may be due to their shared protein domains associated with habitat adaptability in soil. The next branch consisted of other 3 Micrococcales – *Kocuria rhizophila, Brachybacterium faecium* and *Micrococcus luteus*. *Kytococcus sedentarius,* was positioned near *M. luteus* and shared 1229 domains which were mainly associated with ribosome formation (Supplementary File 19a). They were also biologically tightly connected with PPI enrichment p-value of $4.28e^{-11}$ (Supplementary File 19b). The subclade C-IIIb was cluster of Corynebacteriales with all of the *Corynebacterium* strains grouped together. Besides the *Corynebacterium*, C-IIIb also contained *Tsukamurella paurometabola,Segniliparus rotundus*, and *Gordonia polyisoprenivorans* of Corynebacteriales group. The next cluster (C-IV) was a small assembly composed of other Coriobacteriales including *Olsenella uli, Atopobium parvulum*, *Cryptobacterium curtum* and *Slackia heliotrinireducens.*

Clade V (C-V) was a big group which was subdivided into two sub-clusters (C-Va and C-Vb) and was highly branched. The first sub-cluster C-Va composed of Micromonosporales including *Micromonospora aurantiaca, Actinoplanes missouriensis, Salinispora tropica* and *Verrucosispora maris*. Unexpectedly, *Kribbella flavida* was positioned with Micromonosporales in domain based tree whereas, it was found to be with other Propionibacteriales in MLSA tree (Sen et al. 2014). Further investigation revealed 1539 domains were shared among *K. flavida* and Micromonosporales (Supplementary File 20a, 20b). Those domains were mostly allied with secondary metabolites biosynthesis, ribosome formation, bacterial secretion system, homologous recombination, protein export and several metabolic pathways. Thus these domains may help *K. flavida* to maintain a proper lifestyle only in soil. Biological network analysis revealed taut interaction among those domains with PPI enrichment p-value of :< $1.0e^{-16}$ (Supplementary File 20b).

Nocardiopsis dassonvillei, N. alba, Thermomonospora curvata, Thermobifida fusca and *Thermobispora bispora* of Streptosporangiales were placed together in another branch of subclade C-Va. *Stackebrandita nassauensis,* was placed near *S. erythraea* in the following branch which was an exception from Sen et al. (2014). These wo strainsare soil habitatand was found to share 2104 domains among them (Supplementary File 21a). Network analysis revealed considerable connection with PPI enrichment p-value of< $1.0e^{-16}$, (Supplementary File 21b). Moreover, these domains were associated with secondary metabolite production, antibiotic resistance, carbohydrate and amino acid metabolism which are vital aspects for soil habitat. Thus, again this domain based tree depicts its clustering pattern according to the habitat of bacteria rather than their taxonomical facet. The rest of the clustering pattern was allied according to our expectations. C-Vb showed a miscellaneous grouping. For example, Corynebacteriales clustered with the extremophilic stone dwellers *Geodermatophilus obscure*, *Blastococcus saxobsidens* and *Modestobacter multiseptatus* in one branch. In the next branch, *Intrasporangium calvum* and *Nocardioides dokdonensis* were together sharing 1581 domains (Supplementary File 22a, 22b). Two aquatic actinobacteria, *Nakamurella multipartite* and *Microlunatus phosphovorus,* were near each other. The last branch was composed of all plant-associated *Frankia*. The last Clade VI (C-VI) was constituted with *Rubrobacter xylanophilus* (thermal) and *Conexibacter woesei* (soil inhabiting). They were also found together in MLSA tree (Sen et al. 2014). A total of 1283 domains were common between them connected strongly with PPI enrichment p-value of < $1.0e^{-16}$ (Supplementary File 23a, 23b).

Consequently, the domain-based phylogeny grouped actinobacteria strains primarily according to their habitat. Taxonomically, this tree revealed several similarities with that of the previously generated MLSA tree (Sen et al. 2014) supporting the splitting of old Frankiales into Frankiales (containing all *Frankia* strains), Geodermatophilales (*Geodermatophilus, Blastococcus, Modestobacter*), Nakamurellales (*Nakamurella*), and Acidothermales

(*Acidothemus*). However, there were some exceptions where taxonomically unrelated strains grouped together due to the large number of biologically connected shared domains. Those common domains among taxonomically unrelated strains revealed exceptionally strong biological interaction among themselves suggesting their roles in metabolism, signal transduction and niche adaptation.

4.8 Evaluating the biological network of PAS domain

4.8.1 Moderately expressed PAS genes are codon biased

GC3 vs. ENc plots displayed that most of the PAS genes were gathered at the lower right end of the plot alongside the ribosomal protein genes (with some exceptions). A negative correlation between GC3 and ENc was witnessed (i.e. ENc value drops with corresponding upsurge in GC3 value) (Supplementary file 24). Thus the GC3 vs. ENc plot clearly specifies a significantly strong codon bias which may be owing to selection of translational efficiency. This result also proposes that codon usage bias for PAS genes as well as the majority of the genes were affected impartially of the overall base composition. However, correspondence analysis on codon usage indicates no significant variation between the total protein coding genes and PAS genes (Supplementary file 25). The distribution frequency of the CAI values for PAS genes revealed a 49% of the PAS genes having a CAI value between 0.6 and 0.7 and around 48% of the PAS genes having a CAI value of 0.8 or above (rest 3% were between 0.55-0.59). These results may propose that PAS genes are moderately expressed (Supplementary file 26 plots the CAI values of selected actinobacteria representing their niche diversity). Both the GC% and GC3% values of the PAS genes were not statistically dissimilar from the rest of the genome. For all of the studied bacteria, alanine encrypted by GCC was the most preferred amino acid and codon, respectively in both PAS genes (Supplementary file 27) as well as total protein coding genes.

4.8.2 Detection of some PAS genes as HGT

Both Nc/GC3 plot and CoA analysis revealed most of the PAS genes placed with or near the ribosomal genes. However, a few were out-grouped and were placed far from the main cluster. Those genes were diguanylate cyclase

(*Frankia* sp. EUN1f, *Frankia* sp. Eul1c), putative two-component hybrid sensor and regulator (*F. alni* ACN14a), PAS domain S-box proteins (*Frankia* sp. CN3, *Frankia* sp. BMG5.12), PAS fold-4 domain protein, Histidine kinase/PAS fold, PAS sensor protein (*Frankia* sp. EUN1f), hypothetical proteins (*S. coelicolor*, *S. scabiei*, *S. griseus*) and magnesium or manganese-dependent protein phosphatase (*S. avermitilis* MA-4680) (Supplementary file 28). The codon usage pattern of those genes was also different from the rest PAS genes. Correspondence analysis of these out-grouped genes were significantly different from other PAS gens (Mann–Whitney U test, $p \leq 0.001$, Z-score =4.97709). Previous studies suggest that, these differentially characterized genes can be assumed to be horizontally transferred gens. Hence, we may propose that, PAS genes with atypical characters may be considered as horizontally transferred.

4.8.3 Importance of PAS genes being present on lagging strand

The numbers of PAS genes present on leading and lagging strands of each strain under study have been given in Supplementary file 29.A statistically higher amount ($p<0.001$, two tailed test) of PAS genes were found to be present on lagging strand of genomes than on leading strand regardless of the species (Supplementary file 29). Bacteria generally have their vital genes situated on the leading strand of DNA to dodge the head-on collisions that would otherwise happen between replication and transcription mechanisms (Paul et al. 2013). When compared to the co-directional collision occurring including DNA polymerase and RNA polymerase on leading strand, head-on replication transcription encounters have been found to be more mutagenic. Hence, potentially highly expressed house-keeping genes are mostly concerted on the leading strand of DNA (Banerjee et al. 2004). Genes on the lagging strand face more non-synonymous mutations than genes present on the leading strand (Paul et al. 2013). This feature allows an organism to take benefit from the higher rate of mutation by adapting towards precise niche. One hypothesis anticipated that some bacteria with more genes on the lagging strand endorse quicker evolution through orientation-dependent confronts between DNA replication and transcription (Paul et al. 2013). The head-on orientation can be encouraged by a positive selection

mechanism and is obvious from the better rate of convergent evolution in genes on the lagging strand. Thus, the genes on the lagging strand play important roles in adaptive evolution. A simple switch in orientation can accelerate the evolution of genes in a directed way (Paul et al. 2013). PAS genes are mainly present on the lagging strand and may play crucial role in bacterial adaptation through the higher rate of non-synonymous mutation. Fijalkowska et al. (1998) suggested that, lagging strand replication is more precise than leading strand replication, which may be another motive why the majority of these vital PAS genes are located on the lagging strand instead of on the leading strand. However, further research is required to decrypt the prominence of genes existing on lagging strands. Majority (76 %) of the PAS genes whose functions could be dogged and credited to a well-defined COG category were found to be related to signal transduction (COG category-T), whereas, others were poorly characterized [R (10 %), S (14 %)] (Supplementary file 29).

4.8.4 *Tertiary (3-D) modeling of PAS domains*

Homology modeling of the PAS domains exposed their structurally preserved nature (Shah et al. 2013). In its central region, there were 5 antiparallel beta strands linked to each other through loops (Supplementary file 30). Three to four alpha helices were there in the exterior part of the core region, which may confirm that information will specifically transfer to the PAS core region (Scheu et al. 2010). Ramachandan plots obtained from POCHECK server (Laskowski et al. 1993) endorsed the tertiary structures. In the majority of the cases, more than 90 % of the amino acids were found in the allowable region.

4.8.5 Structure based phylogeny

Comparison among PAS structure-based tree (Fig 4.25), PAS domain amino acid sequence based tree (Fig 4.26) and 16S rRNA sequence (Fig 4.27) tree revealed few major topological differences. Analysis of the PAS-domain-structure-based tree presented that members of each clade had a favored association with some co -domains. For instance, all clade-I members had their PAS domain allied with the PAC domain (C terminal motif of PAS). In clade-II, PAS domains interrelated with GAF (cGMP -specific phosphodiesterases) and SpoII E (Stage II sporulation protein E domain). While clade-III PAS domains

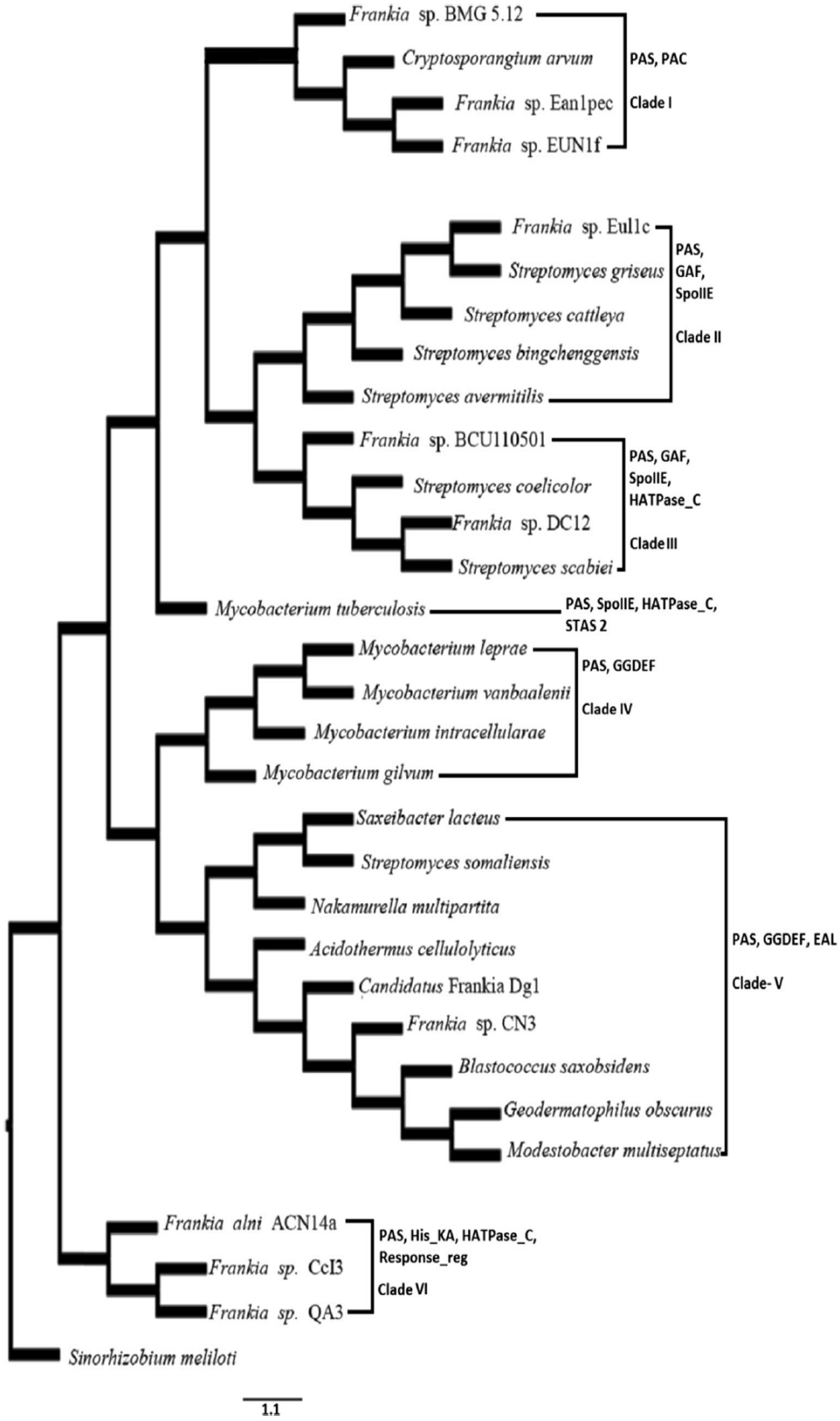


Fig 4.25: A single tree based on the PAS domain structure. *Sinorhizobium meliloti* was used as outgroup. Co-domain association based clusters are clearly shown.

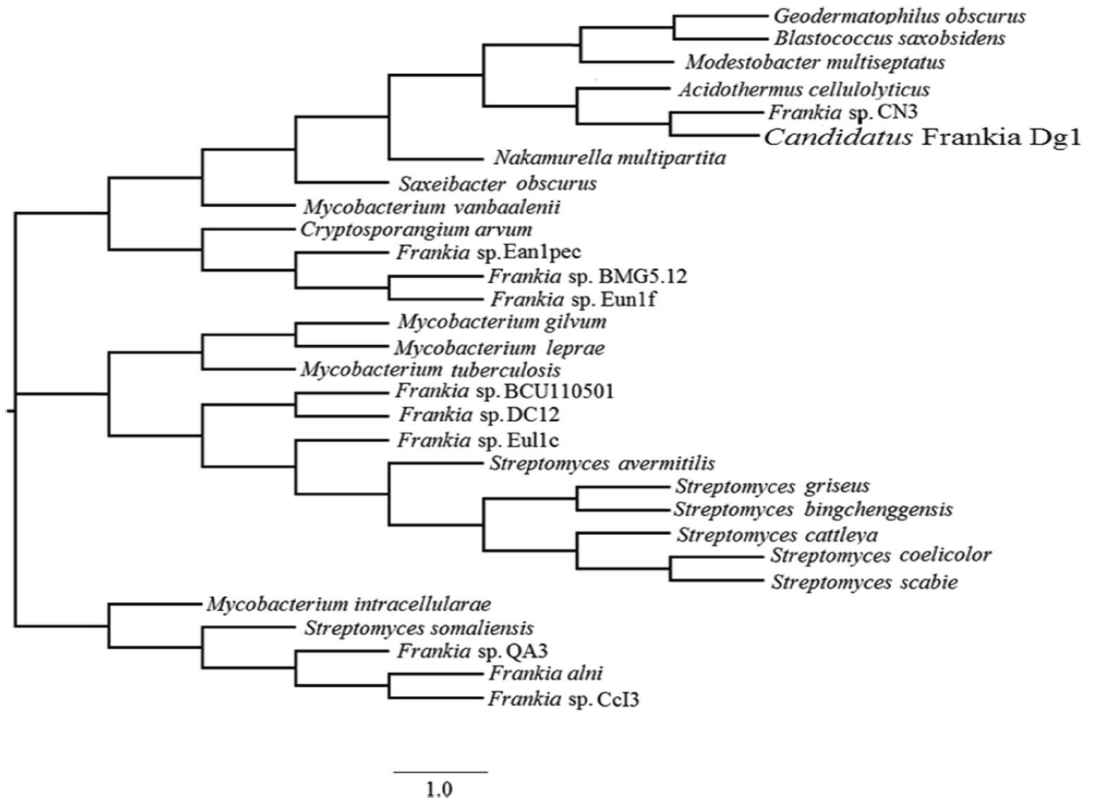


Fig 4.26: PAS domain amino acid sequence based tree

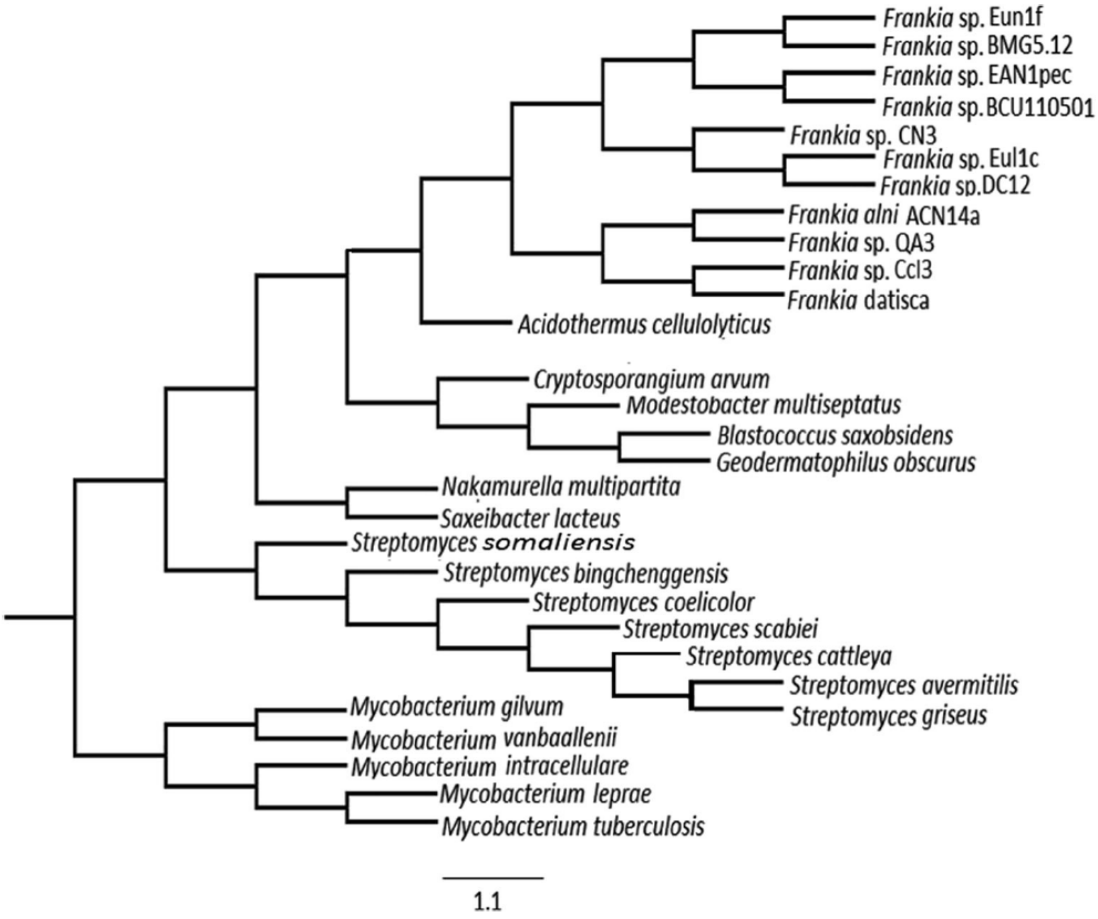


Fig 4.27: 16S rRNA gene sequence based tree

were connected with various co-domains like GAF, SpoII E and HATPase_C (Histidine kinase-like ATPase domain). In clade-IV, GGDEF domains were found to be associated with PAS domains. Members of clade-V were from different orders and their PAS domains were allied with GGDEF and EAL (Glutamate Alanine Leucine domain). This tendency sustained in clade-VI where PAS domains were cooperated with the HisKA, HATPase_C and Response_reg. For *M. tuberculosis* CCDC5079 PAS domain showed exclusive associations with the

SpoIIE, HATPase_c_2, and STAS_2. Thus, bunching pattern of this phylogeny seems to be totally reliant upon the functional communication of PAS domain with its co-domains. Overall, this phylogeny was found to be based on a domain-domain association, regardless of the genus. This supports our hypothesis that, 3-D structure of PAS domain may control the biological network and/or vice-versa. Jonathan and Crosson (2011) reported analogous results and suggested that the structure of the PAS domain gathered according to their cellular localization property and ligand binding property. However, a deeper view of their structure-based phylogenetic tree exposed that PAS domains networked with similar co-domains clustered jointly.

4.8.6 Localization of PAS domains of actinobacteria

The PAS proteins of considered actinobacteria were mostly positioned in the cytoplasm of the cell (Supplementary file 31). Roughly 80.71 % were cytoplasmic, 17.8 % were membrane bound and 1.50 % was extra-cytoplasmic. Proteins with transmembrane regions were membrane bound. The number of extra -cytoplasmic PAS proteins was negligible. Thus, in compare to results of Jonathan and Crosson (2011), we found no significant differences in the cellular localization pattern of PAS protein in studied actinobacteria.

4.8.7 Domain-domain interactions maintaining the biological network

A key feature of proteins is their inter-domain interactions. The PAS domain usually works in a collaborative way with the co-domains of the definite protein sequence (Taylor and Zhulin 1999). An inspection of PAS domain connections with its co-domains stipulated insight on the impact of environmental effects separating the microbes into four biotopic clusters. A

common arrangement of association was revealed in the extremophilic actinobacteria (consisting of *Geodermatophilus, Blastococcus, Modestobacter, Cryptosporangium* and *Saxeibacter* genera) where PAS domains were connected with various co-domains regulating diverse functions. For instance, PAS-MA (Methyl accepting chemotaxis domain) interaction may help in bacterial motility, PAS-GAF-GGDEF, HisKa-PAS-GGDEF-EAL and PAS-GGDEF-EAL may control the microbial signal transduction (Aravind and Ponting 1997), PAS-GAF-PP2C_SIG (protein phosphatase 2C domain) remain associated with formation of endospore and PAS-GAF-PHY (Phytochrome related domain) is responsible for the bacterial phytochrome related activity. Previous researches (Bibikov et al. 2000; Kataoka et al. 1997; David et al. 1988; Stock et al. 1989, Aravind and Ponting 1997) have shown that PAS domains linked with these co-domains can regulate bacterial motility, phytochrome-related activity and other signaling cascade mechanisms. Certain strains of *Frankia* form a N_2-fixing symbiotic relationship with woody actinorhizal plants (Normand et al. 2012). For these *Frankia* strains, majorities (97 %) of PAS domains were found to be allied with either HisKA, HATPase_C (47 %) or with GGDEF, EAL (50 %) and the rest 3 % were associated with SpoIIE, Response_reg, PAC etc. In *Sinorhizobium meliloti* RM FixL gene, PAS was associated with HisKA and HATPase_ C. PAS-HisKA-HATPase_ C association may induce nifA and nifK genes which in turn can trigger the N_2-fixing cascade under hypoxic situation (Gilles and Gonzalez2004). In *Rhizobium etli* too, PAS-GGDEF-EAL association helps in *Rhizobium*-legume symbiosis (Gao et al. 2014). Thus, the connection of PAS with HisKA, HATPase_C or GGDEF and EAL may help in *Frankia* symbiosis too. Transcriptomic data analysis of *Frankia alni* ACN14a (GSE18190) (Alloisio et al. 2010) also exposed that, the expression level of proteins containing PAS-GGDEF-EAL and PAS-HisKA-HATPase_C were raised under symbiotic conditions when compared to free living conditions (Supplementary file 32). Those proteins were mostly in cytoplasmic and present on lagging strand (the reason of being so has already been discussed above). Thus, we may propose that the PAS domain

associated with these co-domains may assist the bacteria in maintaining a proper symbiotic association and in the N_2 fixation process. Further proof on this notion requires experimental evidence beyond *in-silico* results. Among the free living *Frankia* strains, PAS domains were found to be networked with GAF, SpoIIE, Response_reg, GGDEF and EAL. In free living microbes, both GGDEF and EAL dynamically contribute in the secondary messenger pathway and act as equivalent response regulator with sensory PAS domain (Simm et al. 2004; Taylor and Zhulin 1998). The other co-domains may be acting in analogous way as found with other soil bacteria by upholding the cell signaling cascade for important actions like signal transduction, chromophore formation, motility, sporulation etc. (Taylor and Zhulin 1999).

Thus, we hypothesize that *Frankia*, which can exist in both symbiotic and free-living state, utilize PAS-GGDEF-EAL relationship in an alike manner which is very interesting and to be further explored later. For both pathogenic and non-pathogenic *Mycobacterium* and antibiotic producing *Streptomyces* strains, PAS domains were mainly networked with GGDEF, GAF, SpoIIE. In non-pathogenic *Mycobacterium* and soil living *Streptomyces*, the role of the PAS domain was quite identical to free living *Frankia*., PAS domains in the potentially pathogenic *Mycobacterium* strains have been recognized as potential drug target for the treating tuberculosis (Cui et al. 2009; Rickman et al. 2004). PAS has been anticipated to play a crucial role in stress regulation in host-associated *Mycobacterium* too (Jaiswal et al. 2010; Rickman et al. 2004). Thus the importance of tertiary structure of PAS domains in maintaining biological network is revealed.

However, all proposed mechanisms for PAS domains in actinobacteria are still very theoretical and need advance research.

4.9 Novel drug targets identification from tuberculosis strain

4.9.1 Identification of pathogenic and host metabolic pathways

The genomic study of different metabolic pathways present in *Mycobacterium tuberculosis* RGTB423 (MtR) was carried out for the identification of potential therapeutic drug targets. A disciplinary workflow was followed involving several

bioinformatics tools, databases and drug target prioritization parameters with the aim of obtaining proteins which could further serve as receptors for *in silico* drug designing process. Metabolic pathways are known to be vital for maintenance of homeostasis of an organism. Furthermore, targeting a protein involved in a pathway present in both MtR and human may lead to some obvious side effects in the host. Hence, we aimed to determine pathogen specific metabolic pathways in MtR which are totally absent in human. The metabolic pathways of human and MtR were found to be 302 & 117 respectively. Eighty three pathways were obtained as common pathways present in both human and MtR. Detailed pathway studies bring out that a total no. of 34 pathways were present only in the pathogen and absent in host. The essential proteins involved in those pathways were determined to ensure that the proteins we are working on are all vital for survival of the organism and are also involved in pathogen-specific metabolic pathways. To further ensure that, we are targeting only the non-host essential proteins of MtR, a BLASTp analysis was done with the UEMP (against *H. sapeins*). This ultimately gave us 145 MtR proteins which are unique in the mentioned pathogen. These few steps also lowered down the number of workable proteins.

4.9.2 Screening of non-host essential proteins

To reduce the time required for drug testing and development, the addition of gene essentiality information and drug prioritization parameters offers great advantage in the careful selection of candidates for drug discovery pipelines. The "knockout-rate" prediction states that essential genes should be more evolutionarily conserved than are nonessential genes (Jordan et al. 2002). This is because negative purifying selection acting on essential genes is expected to be more stringent than that for nonessential genes, which are more functionally dispensable and redundant. Complying with this idea, Wilson et al. (1977) proposed that essential genes should evolve more slowly than nonessential genes. The matter of the fact is since we have all the essential non-host proteins we are ready with our workable set of proteins.The total no. of essential proteins found in MtR was 598, among which 145 proteins were found to be essential as well as unique

to MtR (i.e. these essential proteins were unique for MtR).

4.9.3 Qualitative characterization

4.9.3.1Molecular weight

We have performed some qualitative analysis (i.e. molecular weight determination, sub-cellular localization, interactome study etc.) on the previously described set of proteins to further refine our final list of drug targets. We found 143 proteins out of 145 which fall within the MW range of 100kDa. A wide range of MW was obtained for example, KEGG ID - MRGA423-03515 had weight of 96.31KDa and the protein having KEGG ID-MRGA423-16805, whose weight was 6.35 KDa had the lowest weight (Supplementary file 33). Proteins with MW<100kDa were selected since, they act as better drug targets because of their more solubility and easier purification rate (Duffield et *al. 2010).*

4.9.3.2 Subcellular localization analysis

Subcellular localization is another important factor while searching for drug targets. Cytoplasmic proteins are reported to be excellent drug targets while, membrane proteins act as vaccine targets (Barh et al. 2011). Hence, we searched for the localization of the aforementioned MtR protein set and selected only cytoplasmic proteins for our further analysis. Subcellular localization analysis of a protein reveals whether the protein is suitable as a drug or a vaccine target. Out of the 143 proteins, 108 were predicted to be cytoplasmic protein and are used for further studies (Supplementary file 33).

4.9.3.3 Interactome analysis

The interactome study reveals the interaction between proteins and gives a confidence score after the analysis of various parameters. The 108 proteins of interest were checked for interaction among them. The combined association score for many such proteins were found to be 0.99. Such observation clearly indicates that the target proteins were interactive ones (Supplementary file 34). Protein–protein functional interactions mediate many essential processes for maintaining the stability of the biological systems. Thus, these interactions plays crucial role in the evolutionary processes of different organisms. In functional study analysis, the most useful networks are those that integrate all types of interactions: stable physical associations, transient binding, substrate chaining,

information relay etc. These functional associations also give an idea about the chokepoint of the organism (Hasan et al. 2006). Our interactome study revealed that chosen MtR proteins are highly interacting with considerable confidence score and they were ranked according to their druggability score.

4.9.3.4 Druggability analysis

Out of 108 proteins, 20 had drug bank entries with drug bank hits ranging from 1-10. Two proteins (KEGG ID-MRGA423-08870 and MRGA423-08330) have highest drug bank hits of 10 and thus may be assumed as most druggable (Supplementary file 35).

4.9.3.5 Ranking the putative therapeutic targets

Drug Target Score (DTS) was calculated by the equation described in materials and methods section which helps to rank up the highly druggable targets. According to the result 6,7-dimethyl-8-ribityllumazine synthase having KEGG ID-MRGA423-08870 showed the highest score followed by DNA-binding response regulator TrcR having KEGG ID-MRGA423_06450 and two component sensor Y transduction protein with KEGG ID MRGA423_03080.

4.10. Validation of the identified drug targets

4.10.1. Receptor Protein Structure

It is the normal convention that, if all of the proteins in the crystal are aligned in an identical way it forms a perfect crystal as all of the proteins will scatter X-rays the same way and the diffraction pattern will show the fine details of crystal. On the other hand, if the proteins in the crystal are all asymmetrical or jagged, due to local flexibility or motion, the diffraction pattern will not contain clear information. Resolution is a direct measure of the details that will be seen when the electron density map is calculated. High-resolution structures, with resolution values of 1 Å or so, are highly ordered and it is easy to see every atom in the electron density map. Lower resolution structures, with resolution of 3 Å or higher, show only the basic shape of the protein chain. As a general rule of thumb, we have more confidence in the location of atoms in structures with resolution values that are small, called "high-resolution structures" (Murshudov et al. 1997). Structures were also screened on the basis of their origin proteins from *Mycobacterium tuberculosis* were only considered here.

Among 108 predicted target proteins 46 showed 100% alignment when BLASTp was performed against PDB Database and all of them were found to have valid by X-Ray cristallographie structure. However, when we checked the resolution for each of PDB structure we found only 21 proteins qualify the >2Å resolution. Among these 21 proteins, 3 proteins were of non *Mycobacterium* origin and were not considered further .These leaves us with 18 workable proteins receptors.

4.10.2. Structural Refinement

Structural refinement was carried out on a knowledge based potential of mean force and stereochemical correction. There was significant energy variation found against many probable receptor proteins. Incidentally, ribose-5-phosphate isomerase B (KEGG ID: MRGA423_15375) has the highest energy variation of -63451.3187 kcal/ mol with a moderate global distance test high accuracy score while, 3-dehydroquinate dehydratase has the lowest energy variations of -2424.6461 kcal/mol. Since we didn't consider any reference structure, the initial and final structures were used for these calculations. Refinement of the structures via KoBaMIN and using embedded ENCAD and KB01 potential force polished the structures that were to be used further. There were significant energy changes in few of the proteins and their confidence score were taken into consideration while selecting them (Table 11).

4.10.3. Ligand selection

Dr. Duke's Phytochemical and Ethnobotanical Databases has chemicals reported for *Aloe vera* and *Withania somnifera*but not against *Oplopanax horridus*. The database has 163 chemicals reported for *W. somnifera* and 313 chemicals for *Aloe vera.* Among 163 chemicals of *W. somnifera* 20 chemicals have various activities reported against it and for 313 chemicals of *Aloe vera* 73 chemicals have reported activities against it. When our desired activity was searched 7 chemicals of *W. somnifera* and 12 of *A. vera* were found suitable. From the literature survey 6 of the *O. horridus* chemicals were chosen as suitable ligand. Further the reported drugs against our protein targets which had repeated DrugBank hits were also considered. This apart, the general prescribed drug against tuberculosis namely Ethambutol and Streptomycin were also used as standard ligands to verify the activities

of our phytochemicals. Based on all this the final list (Table 12) of ligands was completed.

4.10.4. Molecular Docking

AutoDock was employed to detect significant interaction among the target proteins and the selected ligands present in the plants. The ligands selected for this docking experiment were of three different types -the chemicals present in plant sample, the Drug Bank compounds which had maximum hits with our target proteins and the commonly prescribed drugs. The combination displaying the best binding affinity with RMSD (Root Mean Square Deviation) of zero was considered for our analysis. Aloeemodin of *Aloe vera* showed good docking score (Table 13, Fig 4.28). Cardinol and nerolidol of *Oplopanax horridus* also had good activity with some proteins. Cycloartemol and Stigmasterol of *Withania somnifera*s howed good activity (Fig 4.28).

The ligands and the corresponding proteins having significant binding affinity are tabulated (Table 13, Fig 4.28). Apart from good binding affinity with these phytochemicals, few interesting facts came up from the docking results. All the phytochemicals showed better interaction with ribose-5 -phosphate isomerase B which have a PDB ID of 1USL. Among all the phytochemicals, cefoperazone showed the highest interaction in average with all the proteins followed by aloeemodin and stigmasterol. Cefoperazone has drugbank entry of DB01329, it's an approved drug though not reported against Tuberculosis (Fig 4.28). Before going any further it is important to chalk out the general groups of ligand tested here. There are clearly three groups, among them first are the phytochemicals present among the considered plants, second group contains the ligand that has entry in drugbank and third group contains the drugs generally prescribed against tuberculosis. It is also seen that when the average interactions of all the three groups were calculated phytochemicals (present in selected plants) had better interaction than the Drugbank chemicals as well as prescribed medicine. When we considered activity on the basis of plants it is seen that there is not much to pick between them although *Aloe Vera* has slightly better interactions. Aloe emodin of *Aloe vera* showed excellent interactions with all the protein, one such interaction is with

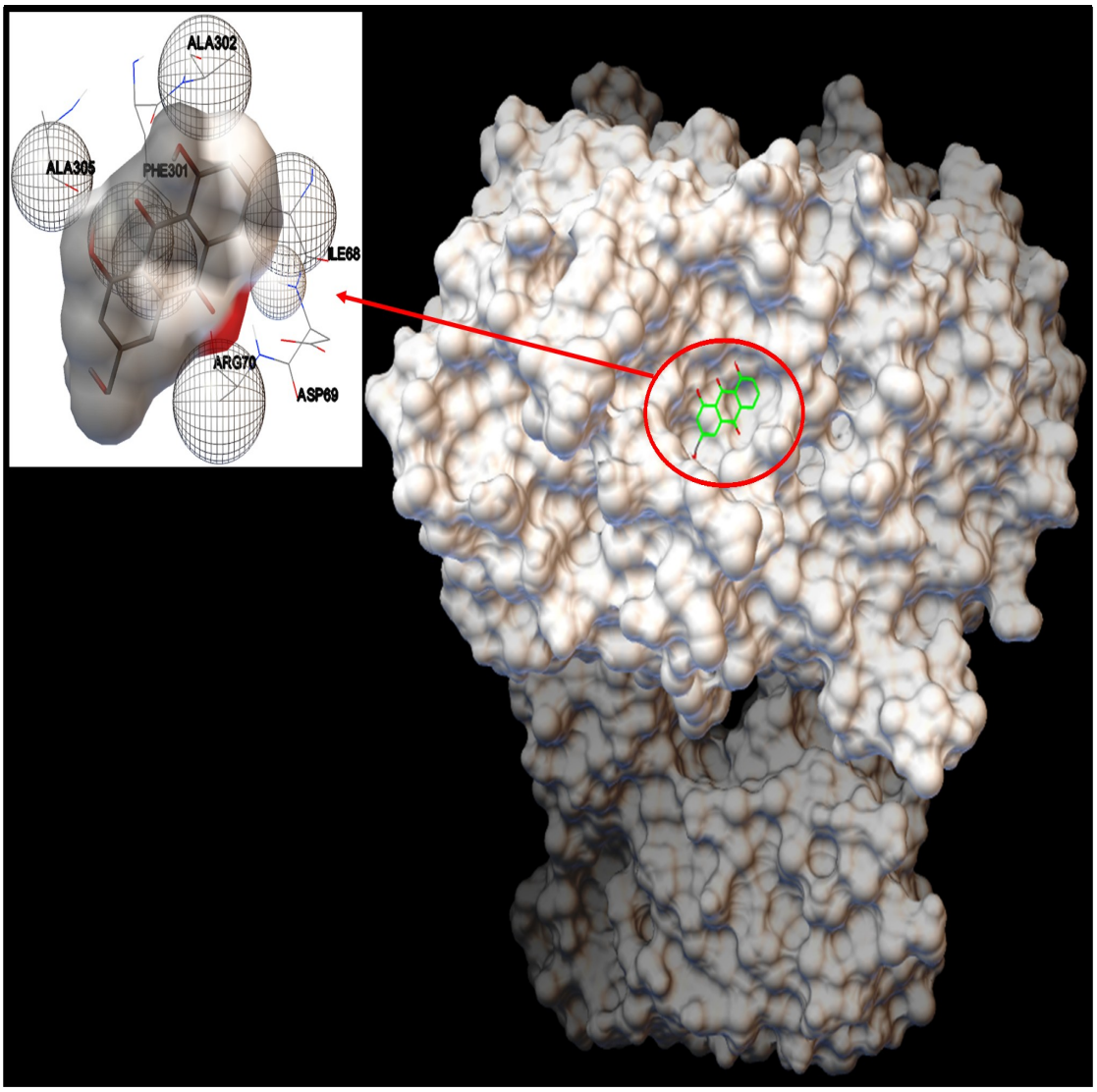


Fig 4.29: Molecular Docking structures: Interaction between 2-isopropylmalate synthase protein having PDB ID 3HPS and aloe-emodin present in *Aloe vera*. On the right is the protein with its molecular surface and coloured entity circled is the ligand (aloeemodin) at its binding site. The top left corner shows the better molecular interaction view of the ligand with the nearby amino acid atoms. The first three alphabets are three letter amino acid codes and the numerical gives the position of amino acids in the protein molecule.

Streptomycin and Glycerol had low value of logP. Nerolidol which had the best individual interaction of -9.5 and a good overall interaction had low solubility with a logD value of 4.32. But Aloe emodin which had the best average interaction among our phytochemicals lies in the ideal logD range. Moreover ligands like Streptomycin, Cefmeatazole, Cefoperazone, Glycerol, Glyoxalate, Stigmasterol failed all the vital Lipinski's criteria (Lipinski 2004) and Veber rule (Veber et al. 2002)

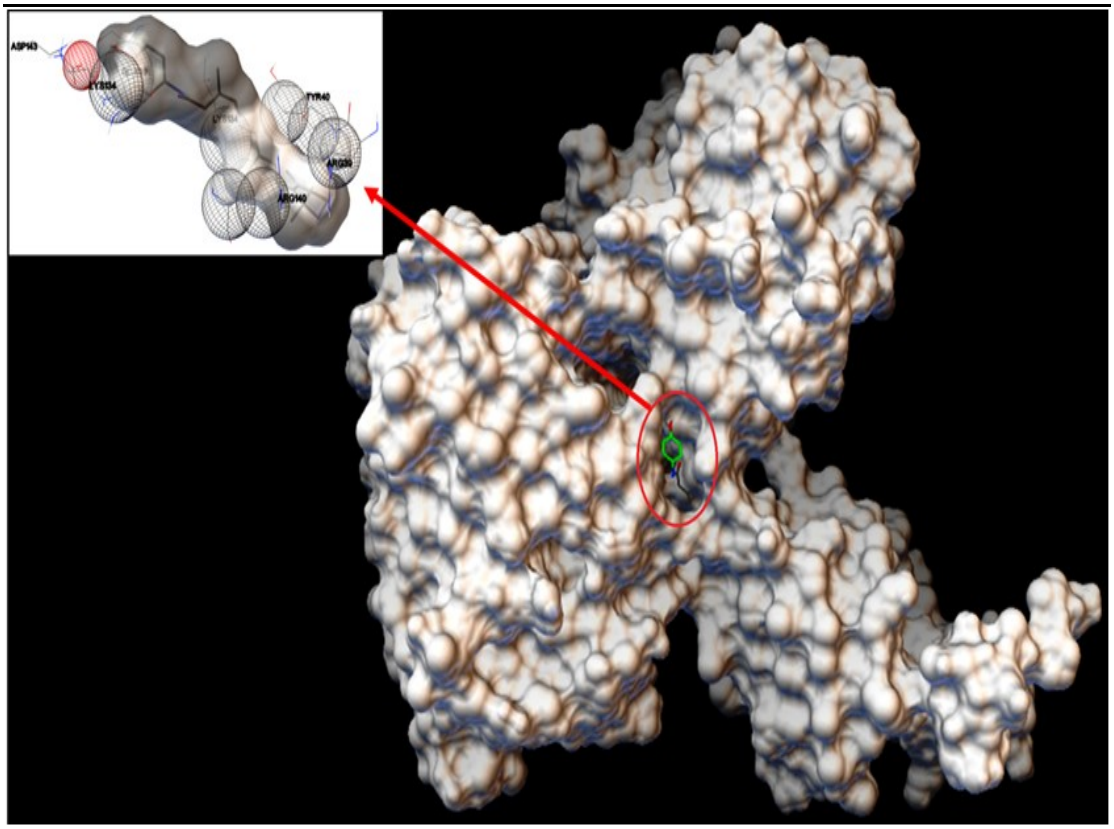


Fig 4.30: Molecular Docking structures: Interaction between ribose-5-phosphate isomerase B protein having PDB ID 1USL and (E)-nerolidol present in *Oplopanax horridus*. On the right is the protein with its molecular surface and coloured entity circled is the ligand {(E)-nerolidol} at its binding site. The top left corner shows the better molecular interaction view of the ligand with the nearby amino acid atoms. The first three alphabets are three letter amino acid codes and the numerical gives the position of amino acids in the protein molecule.

parameters.

4.10.5. Druglikeness

Several druglikeness parameters were explored. To begin with the Lipinski's criteria (https://en.wikipedia.org/wiki/Lipinski's_rule_of_five), Streptomycin which is one of our ligands had more than 5 hydrogen donor moiety and Streptomycin, Cefoperazone and Cefmeatazole more than 10 hydrogen acceptor groups. According to the revised Lipinski's rules (Ghose et al. 1999; Lipinski 2004; Veber et al. 2002) molecular weight of the ligands should lie between 180-500 daltons. From our experimental set Streptomycin and Cefoperazone are overweight and Glycerol and Glyoxalate were found slightly underweight (Table 14). It has also been stated that octanol-water partition coefficient of log P should be between −0.4 to +5.6. Streptomycin, Glycerol and Stigmasterol were found to lie outside the range (Table 14). Following

the Veber rule, the polar surface area of Streptomycin, Cefmeatazole and Cefoperazone has higher Polar surface area and Cefoperazone has excess rotatable bonds. Neo druglikeness parameters like refractiveness revealed Streptomycin, Cefoperazone and Stigmasterol have higher refractivity while Glycerol and Glyoxalate have refractivity lower than ideal range. Similarly, Streptomycin and Stigmasterol had excess number of atoms (Table 14). From the solubility analysis, apart from the log P values, the log S values of Stigmasterol, Glyoxalate, Glycerol lies far from the -5 mol/l mark and thus, were not favorable. Log D values at a pH 7.4 (pH of blood plasma) were calculated and eventually Aloe-emodin and t-Cadinol were found to lie in the ideal range.

4.10.6. Toxicology, carcinogenicity and dosage recommendation

Acute toxicity analysis, dispalying the minimum toxic range of a chemical based on fathead minnow data (Ankley and Villeneuve 2006) was executed. Furthermore, maximum recommended daily dose for human was carried out using machine learning approach. Carcinogenicity analysis on rodents multiple species analysis displayed ethambutol, glycerol, glyoxalateto be active. However, analysis for being oncogenic based on only rat model revealed glycerol and glyoxalate as active. Beta-sitosterol and eugenol were found to be cancer causing when they were tested on mouse model. Mutagenicity analysis revealed Ethambutol and Glycerol as active mutagen (Table 15). Toxicology analysis revealed that, Glycerol possessed active carcinogenicity according to all the three models considered here. Glyoxalate had active carcinogenicity according to rat model and rodents multiple species/sites simultaneously. Ethambutol is also found to be carcinogenic. Moreover, Aloe-Emodin was also found to possess mutagenicity property. However, Cefmeatazole and Cefoperazone do not have much data to do the above analysis. Dosage analysis was available for all the chemicals. It mainly implies that ligands which work in low dose are strong chemicals and phytochemical considered here showed good result in this analysis (Table 15) implicating their effectiveness in treating TB too.