Chapter 1
Introduction

"Arise! Awake! and stop not until the goal is reached."
-Swami Vivekananda

Actinobacteria are Gram positive high guanine+cytosine (G+C) containing bacteria virtually thriving in all ecological niches. Depending upon their habitat, actinobacteria can be classified in 7 different categories (Sen et al. 2014). Figure 1.1 shows the recent classification of Actinobacteria depending upon their niche.

Plant host associated actinobacteria can be symbiotic and pathogenic. Symbiotic microbes (like *Frankia*) generally live in a friendly association with actinorhizal plants (Benson and Silvester 1993). They have the capacity to fix environmental nitrogen (N_2) and provide the nitrogenous compound to their respective host plants. In exchange they get access to host plants' polysaccharides as nutrition

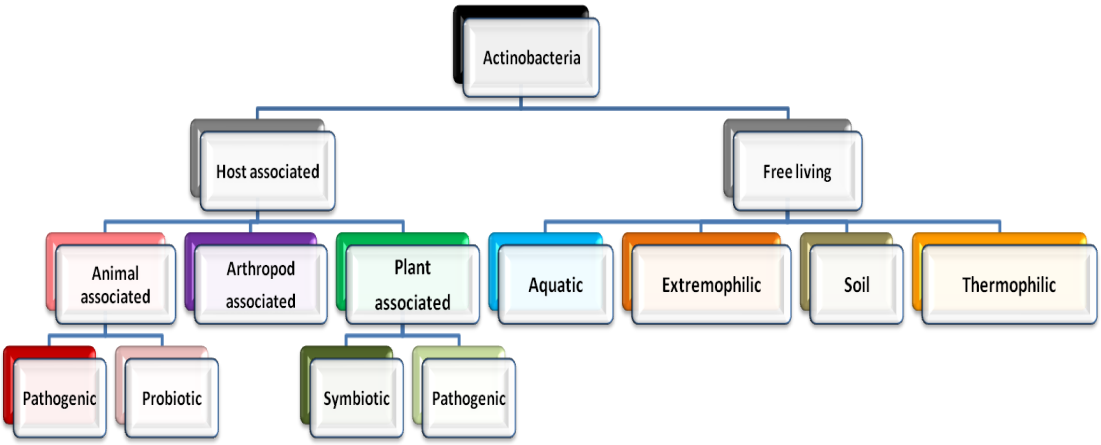

Fig 1.1 Classification of Actinobacteria according to there niche.

(Benson and Silvester 1993). On contrary, few strains of *Rhodococcus* are well known plant pathogens colonizing both exterior and interior aerial parts of plants (Cornelis et al. 2001). Similarly, animal host associated actinobacteria can again be divided into pathogens (*Mycobacterium*, *Propionibacterium* etc.) (Cole et al. 1998, Brüggemann et al. 2004) and probiotics (*Bifidobacterium*) (Roy et al. 2015). Pathogenic genera are popular in medicinal fields due to their potency in causing different diseases (Nordén and Linell 1951). However, these genera also contain some non-pathogenic strains. Genera like *Acidothermus*, *Acidimicrobium* are examples of thermophilic actinobacteria (Sen et al. 2014). On contrary, *Geodermatophilus, Blastococcus, Modestobacter* are extremophilic stone dwellers. Desert sand, artic soil, stones of arid regions and monuments are some predominant habitat for these extremophilic actinobacteria (Sghaier et al. 2016). Studies have reported some very special genomic features among them for instance, UV- resistant capacity, advanced DNA damage repair system etc (Sghaier et al. 2016). *Microccus, Kytococcus, Micromonospora, Verrucosispora, Salinispora* and some strains of *Streptomyces* are under aquatic group (Sen et al. 2014, Hughes et al. 2008). Studies on them have pointed out some crucial genomic approaches adapted by these microbes to sustain in aquatic niche (Sekar et al. 2003, Ghai et al. 2013, Holmfeldt et al. 2009). The quantity of arthropod associated actinobacteria is relative less than other groups. *Bifidobacterium asteroids*, *Tsukamurella paurometabola* and *Nocardiopsis alba* has been identified as arthropod associated actinobacteria (Sen et al. 2014). Last but not the least; the soil saprophyte is probably the most populated group of actinobacteria. These soil habitants (*Streptomyces* mainly) are major source of natural products like antibiotics. Besides, they are also important for their xenobiotic degradation capacity which ultimately provides a deep impact on environment (Barka et al. 2016). Thus, we may say that, each group of actinobacteria is important in their own way and an *in-silico* analysis on their available genomic as well as proteomic sequences may enhance our knowledge of these highly diverse and functionally important microbes. Investigation regarding their codon usage and amino

acid usage prototype may assist us to identify some genomic and proteomic signatures which in long run will improve our understanding towards the chromosomal behavior of these organisms.

In the era of genomics, biological data are being produced at a phenomenal rate. The first complete genomic sequence obtained was that of *Haemophilus influenza* (Fleischmann *et al.*, 1995). Since this major achievement, many genomes have been sequenced and analyzed. These include members of three life domains: Bacteria, Archea and Eukarya. The first genome of a mammalian lactic-acid bacterium, that of *Lactococcus lactis*, a microorganism of great industrial interest, was completed in 2001 (Bolotin *et al.*, 2001) More recently, the genomes of several different microorganisms (Klaenhammer *et al.*, 2002, Kim *et al.*, 2009, Siezen *et al.* 2012*;* Xu *et al.*, 2003; Petrof *et al.*, 2004) have been sequenced. The major driving force for the development of genomics has been the completion of Human Genome Project in 2003. As a result of this surge in data, computers have become indispensable tools in biological research. Such an approach is ideal because of the ease with which computers can handle large quantities of data and probe the complex dynamics observed in nature. Thus, a new discipline, Bioinformatics, has emerged. Bioinformatics is defined as the application of computational techniques to understand and organize the information associated with biological systems and macromolecules. Bioinformatics in the broadest sense includes research in the domains of genome composition, genome expression, proteome analysis, genome and proteome engineering (Perez-Iratxeta *et al.*, 2007). National Institute of Health (NIH), USA, realized the necessity of studying the complexities of human-associated microbes at a greater depth and horizon and accordingly, launched the Human Microbiome Project (http://www.hmpdacc.org/) (HMP). Gargantuan data produced from HMP has provided ample revenues to explore the complex interactions between human host and associated microbes. Emergence of bioinformatics has opened a new window for providing novel understanding into the adaptive policies employed by the bacterial members of the human. Comparative genomics, together with

functional studies, has led to significant advances in this field over the past decade. In the era of genomics, it has been feasible to identify the crucial genetic elements that establish the platform for successful residence of the bacterial populations in the human intestinal environment. Metagenomics is the branch of genomics that performs investigations by direct extraction and cloning of DNA from collection of organisms (Handelsman, 2004). Apart from HMP, Metagenomics of the Human Intestinal Tract (MetaHIT) has been another large-scale sequencing project that has offered scopes to address the complexities of human-associated microbial flora. Genomics coupled with proteomics and transcriptomics are now important high throughput techniques for qualifying and analyzing both gene and protein expression, discovering new gene and protein products, and perhaps these techniques hold the key for revealing the hidden facets of host-microbe interactions.

Advancement of genome sequencing technologies has revolutionized biological research. Plethora of information regarding genome profile can now be mined with the availability of fully sequenced genomes. Study of codon usage patterns of several genes and genomes is a popular technique to characterize and analyze genomic trends from a bioinformatics-based perspective. Codon usage patterns and preferences vary significantly within and between organisms (Grantham *et al.*, 1981; Sharp *et al.*, 1988; Zhou and Li, 2009). The phenomenon of differential codon usage was proposed by Grantham and colleagues (Grantham *et al.*, 1981) in the 'genome hypothesis' theory and it was also stated that codon biases are usually species specific. Codon and amino acid usage profiling of several prokaryotic and eukaryotic forms of life have been successfully accomplished as of now.

Many indices have been proposed to properly elucidate the factors underlying the complex patterns of codon usage and to measure the degree and direction of codon bias (Sharp and Li, 1987). Studies of codon usage can be performed using parameters like GC (Guanine and Cytosine) content of the concerned organism, GC3 (Guanine and Cytosine at the third synonymous codon position) content, relative synonymous codon usage (RSCU) (Peden, 1999), effective number of codons (Nc) (Wright, 1990) and frequency of optimal codons (Fop)

(Ikemura, 1981). Codon adaptation index (CAI) has been proposed as an efficient tool in analyzing the patterns of codon usage within a gene relative to a reference set of genes (usually ribosomal protein genes) (Sharp and Li, 1987). This index has been shown to correlate significantly with mRNA expression levels (Ikemura, 1981) and has been used to predict sets of highly expressed genes in various organisms (Sharp and Li, 1987; dos Reis *et al.*, 2003; Martin-Galiano *et al.*, 2004; Wu *et al.*, 2005).

Availability of completely sequenced bacterial genomes now made it possible to explore the complex riddles of codon and amino acid usage of these organisms and properly address the interplay of various factors that contribute to varying traits.

The rapid progress of genome sequencing has opened the flood gates to explore gamut of data pertaining to genome dynamics and genome complexities in both prokaryotic and eukaryotic life forms. Comparative genomics based analysis attempts to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes (Lukjancenko *et al.*, 2012). Comparative genomics based approaches promise to reveal genetic variation among concerned sets of genomes. Strategic grouping of the genes into functional groups or families has been a smart way of extracting meaningful information (Zakham *et al*. 2012). Such grouping scheme is based on protein sequence similarity, as this approximately predicts conservation of gene function.

Systematic characterization and profiling of the core genome and pan-genome of organisms of interest has been an efficient technique to elucidate the puzzles of speciation and genomic variations. Core genome is referred to as the conserved pool of genes shared by the strains of a concerned species of interest (Medini *et al.*, 2005). Dispensable genome reflects the pool of genes present in some strains of a species but absent in others (Medini *et al.*, 2005). Pan-genome is commonly referred to as the collective repertoire of the core genes, the dispensable set of genes and the unique genes present only in a single strain (Rouli *et al*. 2015). The essence of a species, in terms of its fundamental biological processes and derived traits from a common ancestor, is linked to the core genome.

Rapid progress of HMP and MetaHIT has opened the revenues to explore the features and interaction patterns of inherent bacterial members of human. Comparative investigation of the bacterial genomes representing human microflora, aided by a comprehensive profiling of the pan genome, core genome and the unique gene sets, promise to bestow extensive information pertaining to the adaptive policies employed for successful residence in human.

Protein secretion in bacteria plays an imperative role in communication and cross-talk with other bacterial communities and also with host niche. Exploring the intricacies of bacterial communication and signaling with surrounding host environment has been a challenging chore of present day biological research. Secretomes have been defined as the complete set of proteins secreted by a cell (Ranganathan and Garg, 2009) and are associated with a broad range of functions and critical biological processes, such as cell-to-cell communication and cross-talks, cell migration, and most inevitably virulence and potential infective strategies in disease mechanism (Tjalsma *et al.*, 2004). The signal peptide part of the secreted protein, which is generally composed of around thirty amino acid residues, transports the newly synthesized protein to the protein-conducting SecE and SecY channels associated with the plasma membrane (Leversen *et al.*, 2009). Signal peptides in most cases are reported to possess three domains: a positively charged n-terminus (n-region), a stretch of hydrophobic residues (H-region), and a region of mostly small uncharged residues containing a characteristic cleavage site recognized by a specific signal peptidase (SPase) (von Heijne, 1984; von Heijne, 1989; von Heijne, 1990). It is this characteristic site that holds the key in cleavage of a secretory protein by either of the two SPases, Type I or Type II. Various types of signal peptides are reported in bacterial systems among which secretory signal peptides (Sec type), Twin arginine signal peptides (TAT type), lipoprotein signal peptides (Lipo type), pseudopilin-like signal peptides, and bactericin and pheromone type signal peptides are most prevalent (Tjalsma *et al.*, 2004). However, mainly the first three types of signal peptides (i.e., Sec type, TAT type, and Lipo type) are common in gram-positive bacteria.

Gram-negative bacteria execute protein secretion employing specialized secretion machineries such as type I (ABC transporters), type III (flagellar-type), and type IV (conjugation related) secretion systems (Papanikou *et al.*, 2007; Saier, 2006). Sec type and Tat type signal peptides are cleaved by Type I SPase, whereas Lipo type signal peptides are cleaved by Type II SPase (Storf *et al.*, 2010).

Tremendous advancement in genome sequencing technology has yielded complete genome sequences of a broad range of bacterial population. Automated prediction of secretomes has generated a lot of interest. Prediction of the signal peptide-containing genes, along with their cleavage sites in the completely sequenced bacterial genomes, have been achieved by employing various algorithms such as Hidden Markov Model (HMM), Neural Network (NN) (Bendtsen *et al.*, 2004), and Support Vector Machines (SVM) (Vert, 2002). There have been various web-based servers that employ these algorithms and use specialized programs to predict the secretomes accurately in a given genome. Some of the frequently accessed programs include Signal P, Signal-CF, SIGCLEAVE, Predisi, SPEPLip, SecretomeP and Phobius.

Mammoth genomic data produced from HMP and MetaHIT has provided immense scope to unravel the complex interactive strategies employed by actinobacteria. Study of codon and amino usage patterns, expression behavior, and functional classification of the predicted secretomes among the bacterial masses might confer fruitful information pertaining to their acclimatization in their respective niche.

Molecular evolution of secretory proteins of actinobacteria is another important aspect that demands to be investigated to resolve the mysteries of bacterial adaptation and co-evolution. A reliable index of genetic drift over evolutionary time is the ratio of Ka (non-synonymous substitutions per site) to Ks (synonymous substitutions per site) for a large set of orthologous genes, based on comparisons of related species. The terms Ka/Ks and dN/dS are often used interchangeably. But for the computational purpose, the ratio of non-synonymous (Ka) to synonymous (Ks) nucleotide substitution rates is frequently used as an indicator of selective pressures on protein-coding genes. Ka/Ks ratio reflects the rate of adaptive evolution against the

background rate (Hurst, 2002). This parameter has been widely studied in the analysis of adaptive molecular evolution, and is regarded as a general method of measuring the rate of sequence evolution. To study the impact of selective pressures on the pattern of genetic divergence, it is necessary to find out the pairwise ratio of Ka/Ks between the orthologous gene pairs. Different regions of a single gene can be exposed to varying selective pressures (Hurst and Pal, 2001). In these cases, calculating Ka/Ks over the entire length of the gene does not provide a detailed picture of the evolutionary constraints associated with the gene. Hence, a finer analysis of the Ka/Ks ratio by using sliding windows of different sizes provides minute details of the selective constraints acting on specific positions of a gene segment. Proper assessment of evolutionary signatures of secreted protein components in pathogenic actinobacteria strains provide meaningful know-how pertaining to complex communication tactics and adaptive policies opted by the microbes for successful residence in host environment.

Human beings are susceptible to various microbial pathogens that severely affect their health. Besides being exposed to external pathogens, humans harbor wide range of bacterial communities (members of human microflora) associated with manifestation of severe diseases (Guinane and Cotter, 2013). Investigations of the human genome have led to tremendous advances in the field of biomedical science and drug discovery. Disease etiologies involve the interaction of the human body, external environment and the pathogenic microbes that infect human beings. Proper identification of drug targets in any pathogenic organism is the most vital step in drug discovery process. Rapid progress in genome sequencing technologies and advances in bioinformatics and cheminformatics based research domains have provided a massive scope for enhancement of drug discovery technologies. Availability of genomes of both the host and concerned pathogen provides a platform for subtractive genomics based drug target identification in concerned pathogen (Allsop, 1998; Stumm *et al*., 2002). Subtractive genomics based approach involves subtraction of host genome from the pathogen while screening the tentative targets in the pathogen. Such a

methodology ensures that the drugs, targeted against the essential gene components present solely in the pathogen, do not interact with the human genes. Computational tools have made it easier to filter the unique essential genes in concerned pathogens that are associated with the robustness and viability of the infective organism (Damte *et al*., 2013; Amineni *et al*., 2010).

This present work has been envisaged and developed at the Bioinformatics Facility, Department of Botany, University of North Bengal. The main research motive of this work was to investigate various genomic and proteomic signatures or traits of actinobacteria in totality thriving among different niche and unravel the enigma of complex interactions that these bacterial members exhibit with their corresponding environments. Hence, the proposed objectives of this work were

1. Exploring various metabolic pathway related genes

2. Analyzing variation in energy cost with expression level of proteins

3. Comparative genomics of selected Actinobacteria

4.Comprehensive secretome analysis among selected *Mycobacterium*

5. Detailed secretome analysis among selected Geodermatophilaceae

6. Profiling of Carbohydrate-Active enzymes (CAZymes) of selected *Frankia*

7. Domain based phylogenetic tree generation

8. Biological network analysis of Per-Arnt-Sim (PAS) domains

9. Identification of potential drug targets from *Mycobacterium tuberculosis*

10. Validation of the identified drug targets

Comprehensive investigation of codon usage signatures of various actinobacterial genera residing in diverse niche would be carried out extensively. Our approach would aid in proper explanation of the major forces that influence the codon usage patterns in selected microbes. We will compare the codon usage prototype of the whole organism with the pattern of some metabolic pathway related genes. This analysis will reveal an overall view on codon usage signature of the investigated organisms and how this codon usage is helping the microbes in their niche adaptation. We plan to

explore the complex amino acid usage behavior of the actinobacterial members that dwell in different niche and identify the probable determinants that govern the observed patterns. Energy cost analysis among differentially expressed proteins may reveal some interesting facts associated to the niche adaptive nature of studied microbes.

Comparative genomic analysis of various crucial bacterial genera of Actinobacteria would be performed. Specific and characteristic genomic and proteomic signatures of several concerned bacterial genera would also be plowed into. The pan– and core genomic traits of every concerned genus would be scrutinized with a motive to elucidate the adaptive strategies of the microbes in respective niche. Proper know-how of the core genomic architecture of various actinobacterial members would provide resource for a better understanding of the 'acclimatization policy' adopted by these groups of microorganisms.

In silico prediction and characterization of secretomes, the complete set of secreted proteins, among pathogenic and non-pathogenic *Mycobacterium* strain as well as stone dwelling Geodermatophilaceae group would be executed. Secretomes are believed to execute several crucial 'remote-control' functions associated with cellular communication and cross-talks. Thus, proper identification and know-how of the secretome sets in selected organisms might provide a better elucidation of the host-microbe interactions and adaptive tactics employed by microbes. Amino acid usage tendencies of the secretomes would also be investigated to explore any bias in usage of amino acids among these special sets of proteins. Functional characterization of the secretomes, based on analysis of Clusters of Orthologous Groups (COG) category, would be accomplished to revisit the functional implication of the secretomes. We also intend to compare and analyze the amino acid biosynthetic cost of the secretomes, with respect to the non-secretomes. Evolutionary signatures of the secretomes, with respect to the non-secretome sets, would be investigated with an intention to assess the forces of selection that might be operative on the secretory machinery of the bacterial systems.

We also plan to execute extensive profiling of the special sets of enzymes called Carbohydrate-Active enZymes

(CAZymes) that are prevalent in the plant associated microbes and are associated with the breakdown of complex carbohydrate moieties and glycoconjugates. Microbial CAZymes are able to break down complex plant polysaccharides into simpler form providing nutrients to the bacteria. The carbohydrate components also supply nutritional sources for plant associated organisms to thrive and vibe in the plant. Thus, proper analysis of the CAZymes in selected Frankia might render crucial information regarding the metabolic apparatus of these organisms and their close concord with plant host.

Protein domains are semi-autonomous parts of proteins with their own structure and function. One domain may remain present in several proteins and one protein may contain one or more than one protein domain. Thus a phylogenetic tree depending upon information available regarding the presence and absence of protein domains may provide a deeper view regarding the niche adaptive nature of Actinobacteria. We will select 100 Actinobacteria from diverse niche and perform the analysis to reveal the effect of protein domains in bacterial niche adaptation.

Per-Arnt-Sim (PAS) domains are one of most crucial domains among living organisms. Experiments with these domains have revealed their functionality in signal transduction, stress control, phytochrome activity, tactic movement and niche adaptation. PAS domains are also annotated in all actinobacterial whole genome sequences currently available. However, there are a few works depicting the role of PAS domain among Actinobacteria. Here, we will focus on the impact of tertiary structure along with their co-domain association in different niche associated bacteria. This study may reveal some interesting features of PAS domain and its tertiary structure in biological network regulation.

Tuberculosis is a global epidemic killing a person in roughly every 18 seconds and has become one of the major causes of death worldwide (https://www.tballiance.org/why-new-tb-drugs/global-pandemic). : *Mycobacterium tuberculosis* is the causative agent of Tuberculosis. *M. tuberculosis* RGTB423 strain was isolated and sequenced from India. Our aim here is to search for an alternative herbal medicine to counter the rage of this highly infectious disease. The

workflow planned in this regard is to do a comparative pathway analysis of the host and causative organism. The proteins present in these pathways will be characterized and screened so that they can be used as targets for in silico drug designing tests. The medicinal plants to be considered in this study are *Withania somnifera*, *Oplopanax horridus* and *Aloe vera*. The phytochemicals available in these plants will be used as working ligands for this investigation. Molecular docking of these phytochemicals will be performed with the proteins identified as potential drug targets. Few ADMET parameters and *in silico* toxicology data will also be checked for the ligands to confirm their druglikeness characters. The above tests including molecular docking will be carried out on a few prescribed drugs and a few DrugBank chemicals which will be considered as standard for our exotic phytochemicals.