Chapter 1: Introduction

1.1 Microbial diversity

Microbes comprise the majority of the diversity of species on earth. They are everywhere, occupying every chunk of the planet earth, maintaining its geochemistry and cycling of elements, and managing degradation of wastes. For example, there are about 40 million bacterial cells in one gram of soil; one microliter surface seawater has been estimated to contain thousands of different bacteria, archaea, and tens of thousands of different viruses; a millilitre of fresh water contains a million bacterial cells[1, 2]. The realization that most microbes are hitherto uncultured forced microbiologists to question their belief that the microbial world had been conquered as there is paucity of information about the range of metabolic and organism diversity in any given environment[3]. Given the present state of knowledge, it is conceived that the microbes inhabiting varied ecosystems contain a plethora of genetic resources that if explored will open up new avenues in all areas of industrial, agricultural, and biomedical research. The advancements in molecular biology techniques and bioinformatics tools in the second half of the twentieth century have paved a way to better comprehend the microbial ecology, physiology, genetics and their interactions with the ecosystems they inhabit. Several genome-based molecular methods have been developed to explore the untapped genetic resources and overcome the difficulties and limitations associated with culture based techniques. One such molecular tool that is indispensable in order to retrieve the untapped genetic resources is metagenomics.

1.2 Metagenomics

Also known by other names such as "environmental genomics" or "community genomics" or "microbial ecogenomics", metagenomics is the study of collective microbial genomes isolated directly from environmental samples and does not rely on cultivation or prior knowledge of the microbial communities[4]. The term is derived from the statistical concept of meta-analysis, the process of statistically combining separate analyses and genomics, the comprehensive analysis of an organism's genetic material[5]. This is a rapidly growing field of research that aims at studying all the genomes (metagenome) collectively to

understand the true diversity of microbes, their functions, interactions and evolution, in environments such as soil, water, digestive system of humans and animals, or the rumen of ruminants.

Environmental metagenomic libraries have proved to be a great resource for novel microbial agents, enzymes and antibiotics with potential applications in biomedical and industrial biotechnology[4, 5]. The construction of metagenomic library involves the following steps: (1) isolation of total community DNA from an environmental sample; (2) cloning of random DNA fragments (prepared by a partial restriction digestion of the metagenome) into a suitable vector; and (3) transforming the clones into a heterologous host and screening for positive clones (Fig. 1.1).

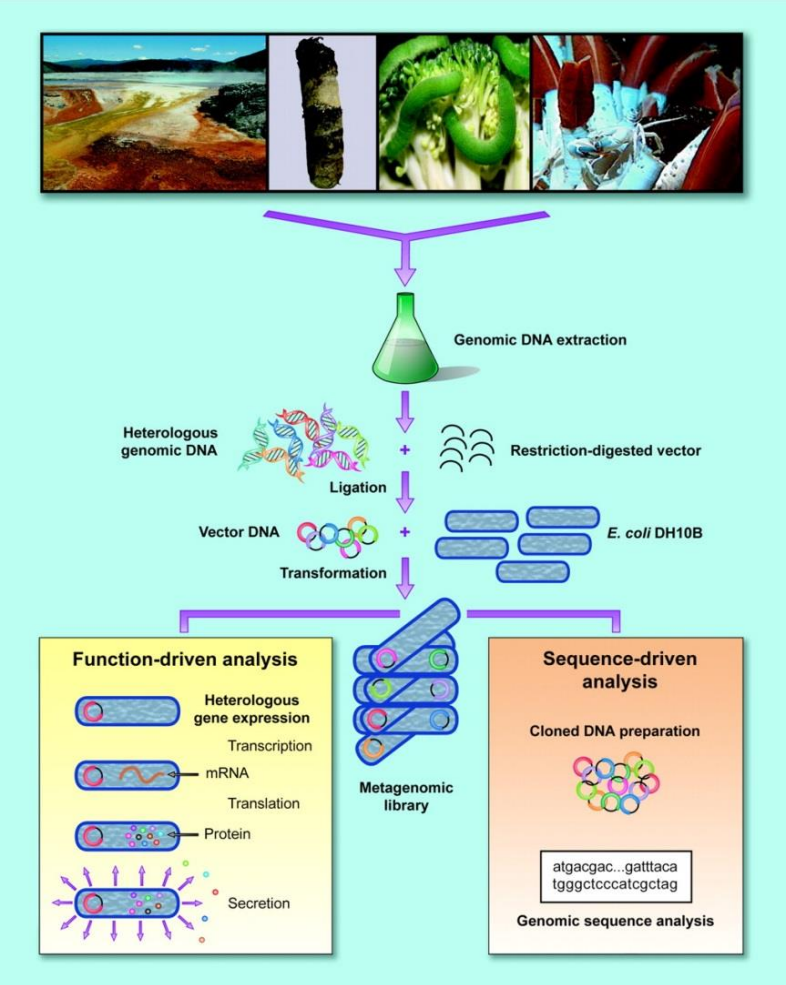

Fig. 1.1 Schematic representation of construction of metagenomic libraries from environmental samples (Handelsman[3])

Metagenomic libraries containing smaller DNA fragments (2–3 kb) provide better coverage of the metagenome of any given environment than those with

larger fragments. Essentially, metagenomics does not include methods that investigate only PCR-amplified genes (e.g., DNA fingerprinting) as they do not provide information on genetic diversity beyond the genes that are being amplified[6].

The metagenomic libraries can be screened using either the sequence-driven metagenomic analysis or by functionally screening the expressed phenotypes.

1.2.1 Sequence-based approach

Sequence-based approach of metagenomics involves massive high-throughput sequencing of the metagenomic clones and elucidates many important genomic features such as redundancy of functions in a community, genomic organizations, and traits that are acquired from distinctly related taxa through horizontal gene transfers[3]. The significance of this approach is that it is independent of the expression of cloned genes in heterologous hosts.

The assessment of microbial diversity in any particular environment is done by sequence-driven metagenomic analysis. It makes use of various techniques viz. 16S rRNA gene analysis for identifying member community, substrate utilization for measuring metabolic diversity, DNA-DNA re-association kinetics to view genetic complexity, temperature gradient gel electrophoresis (TGGE)[7] etc.

1.2.2 Function-based approach

In function-based metagenomic analysis (functional metagenomics), the metagenomic libraries are screened for the expression of a desired phenotype on a specific medium. Recent progress in mining the rich genetic resource of hitherto uncultured microbes has led to the discovery of new genes, enzymes, and natural products. For example, novel antibiotics (e.g., Turbomycin, Fasamycin A and B, Terragine, Indirubin), microbial enzymes (e.g., cellulases, proteases, lipases, amylases), and proteins (e.g., antiporters) have been identified in soil metagenomic libraries[5, 8, 9]. Using metagenomics for exploiting the whole microbiome (genomic material of all microbes inhabiting a specific niche) of a given environmental sample has met a lot of success through path-breaking discoveries as far as the industrial sector is concerned. Through metagenomics, the quest for novel

biocatalysts has yielded positive results with novel enzymes being unlocked from the genetically untapped resources that find applications in varied fields (Fig. 1.2).

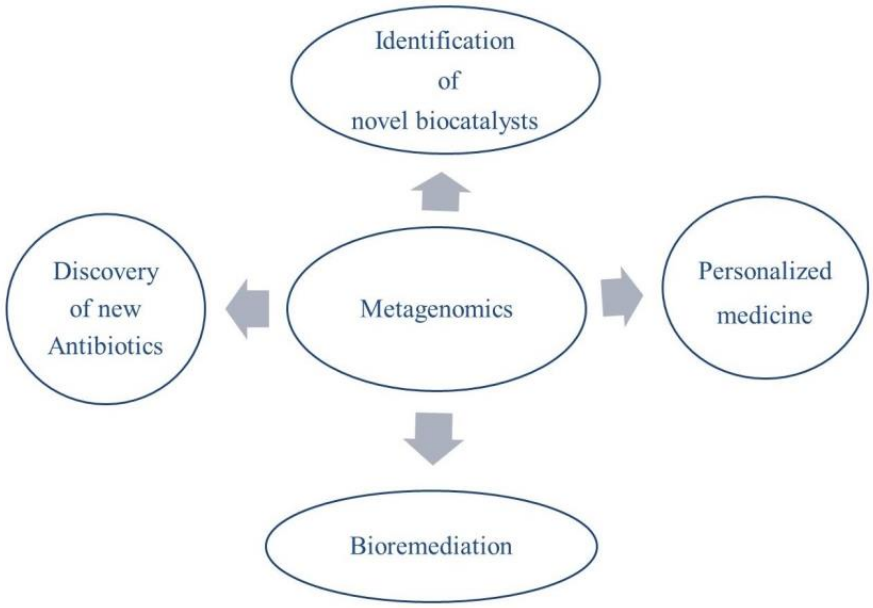

Fig. 1.2 Industrial applications of metagenomics

The economic production of antibiotics and biosurfactants using metagenomics has given an alternative for addressing the issues of drug resistance and oil leakage, respectively. Furthermore, metagenomics has been used for the breakdown of synthetic compounds, crucial to protect the environment from pollution. Metagenomics thus provides a means to look into the structural and functional genomics of microbial diversity and paves a way to discover novel genes and operons encoding enzymes and bioactive molecules of industrial and biomedical importance with better properties[8].

The function based approach of metagenomics has also proven to be vital in the discovery of optimal catalysts and enzymes for a broad spectrum of biotechnological applications. New genetic information of various industrial enzymes such as cellulases, esterases/lipases, amylases, proteases, and others has been produced by metagenomic approaches[8].

Therefore the sequence-driven approach of metagenomics allows us to address the question "who is there?" and the functional metagenomics investigates "what are they doing?" and "how are they doing it?"

1.2.3 Why metagenomics study?

A metagenomic study is specifically taken up for the following reasons:

- Assessing microbial diversity using 16S rRNA and other phylogenetically informative genes; microbial diversity patterns for monitoring and predicting environmental conditions and changes.
- Exploring genes/operons for desirable enzymes and other metabolites (e.g., proteases, cellulases, lipases, chitinases, antibiotics etc.) for potential industrial or biomedical applications.
- To keep a track of variations within genes for the important enzymes which might help to identify and design the optimal catalysts.
- Discovering novel metabolic pathways. A comprehensive understanding may lead to directed approaches towards designing culture media for the growth of previously uncultured microbes.
- Exploring genes that predominate in a given environment as compared to others. Recognition of "marker" genes in an environment can assist in developing methods to understand community responses, interactions, and evolution.

1.3 Profiling of microbial diversity

It is a well-known fact that with the conventional laboratory cultivation techniques only about 1% of the bacterial diversity has been explored in any given environment and that the genomes of the total microbiota found in nature (metagenome) contain vastly more genetic information than the culturable subset. In metagenomics, the DNA is extracted directly from an environmental sample. Factors such as genome size, genome copy number, within-species heterogeneity, and the relative abundance of the species, as well as biases in DNA extraction and sequencing, determine the number of reads that are derived from a given species. Thus, the culture-independent molecular techniques require high quality DNA as template and effective DNA extraction methods to ensure that the extracted DNA represents the total community gene pool. Any variation in the ability to lyse cells, microbial structures and the complexity of inhibitory substances such as humic acids, can lead to biases in microbial diversity studies.

Environmental genomics projects based on cultivation-independent approach originated from Norm Pace's group's study of natural microbial

populations[10, 11]. They used ribosomal RNA (rRNA) sequences to define and enumerate the components of mixed, natural populations, and shotgun-cloned 16S rRNA genes using DNA purified from natural samples. This laid the foundation for direct cloning of genes from environmental samples bypassing the need to culture individual strains of microbes. One of the first large-scale microbial metagenomic projects was undertaken by Venter and colleagues to examine microbial populations within surface waters of the Sargasso Sea. They estimated that over 1,800 genomic species were in the sample, including 148 novel bacterial phylotypes[12]. Advances in DNA sequencing technology, construction of improved cloning vectors and streamlined cloning techniques facilitated the recovery and sequencing of large DNA inserts from microorganisms, and thus increasing the feasibility of metagenomic projects. In the early 1990s, only a limited number of groups were able to sequence DNA up to 100,000 bases and that too at extremely high costs. Celera Genomics and the Human Genome Project (United States Department of Energy) took an initiative to sequence the human genome and this inspired scientists and engineers to design automated techniques that not only speed up the process of DNA sequencing but also substantially lower its cost. Much advancement has been made to the 1st generation sequencing techniques viz. Maxam-Gilbert's sequencing[13] and Sanger's sequencing[14]. Now 2nd generation sequencing techniques (Next-generation sequencing) such as Roche/454 pyrosequencing (http://www.454.com), Illumina (Solexa) (http://www.illumina.com), SOLiD[15], Paired-end or mate-pairs sequencing[16, 17] and 3rd generation sequencing techniques (Next-Next-generation sequencing) such as HeliScope[18], SMRT sequencing[19] (http://www.pacificbiosciences.com), Ion Torrent, and Nanopore sequencing[20] are in use that generate long reads up to several thousand base pairs. DNA sequencing is now done routinely and many laboratories sequence 100 million bases or more every year.

Metagenomics has become most direct and unbiased means to interrogate the functional potential of microbial communities[21]. The extraction of total community DNA from environmental samples, polymerase chain reaction (PCR) amplification of SSU rRNA genes using universal primers followed by cloning and sequencing of these genes have now become commonplace and often the first steps in studying the microbiology of an environmental sample of interest[22].

1.4 Enzymes

Enzymes are biological macro-molecules that are responsible for thousands of the chemical inter-conversions that sustain life. Enzymes are highly selective catalysts that accelerate both the rate and the specificity of all metabolic reactions, from the digestion of food to the synthesis of DNA. Most enzymes are proteins, each adopting a specific three-dimensional structure for its activity and may employ co-factors and prosthetic groups to assist in catalysis. Almost all biochemical reactions in a cell need enzymes in order to occur at optimal rates to support life. Enzymes work by lowering the activation energy, thus increasing the rate of the reaction. As a result, products are formed faster and reactions reach their equilibrium state more rapidly. As with all catalysts, enzymes are neither consumed by the reactions they catalyse, nor do they alter the equilibrium of these reactions. Enzymes are key to all biosynthetic pathways in the production of natural metabolites in microbes. They have the unique ability of producing chiral molecules as intermediates in chemical and drug synthesis, which is usually not possible with traditional chemical methods. They are also vital in various other industries such as pulp and paper, textile, detergent, food and nutrition, dairy etc. Enzymes with highly efficient reaction rates, unique pH optima, thermostability, solvent tolerance, substrate specificity, halotolerance are desirable as far as industrial applications are concerned. One such enzyme that is industrially very important and vital to the production of second generation biofuels and meeting the global energy demand is cellulase.

1.4.1 Cellulase

Cellulases act on cellulose, a linear polymer of β-linked glucose molecules that are organized into higher order fibrillar structures. Cellulose is the most abundant biopolymer[23]. It is the major polysaccharide compound in plant biomass and the most abundant source of renewable energy[24]. Previous studies in natural cellulose hydrolysis have revealed many cellulolytic microorganisms and their complex cellulases[23, 25, 26]. Cellulases are inducible enzymes synthesized by a large diversity of microorganisms including both fungi and bacteria during their growth on cellulosic materials[27, 28]. Among them, the genera of *Clostridium*, *Cellulomonas*, *Thermomonospora*, *Trichoderma*, and *Aspergillus* are the most extensively studied cellulase producers[29, 30, 31]. Bacteria, by virtue of their immense

natural diversity and quick growth, have the capability to produce highly thermostable, alkali-stable enzyme complement and may serve as highly potent sources of industrially important enzymes[32].

The enzymatic degradation of cellulose in nature requires the coordinated action of three major types of cellulases: endoglucanases (EC 3.2.1.4) that act randomly on internal amorphous cellulose fibrils; exoglucanases, including cellodextrinases (EC 3.2.1.74), and cellobiohydrolases (EC 3.2.1.91) that sequentially release sugar units from both ends of cellulose chains; and β-glucosidases (EC 3.2.1.21) that convert glucose dimers and oligomers into monomers for instant energy[33, 34, 35]. Cellulases are of significant industrial importance finding use in textile processing, kraft pulp bleaching, preparation of human and animal foods, brewing and wine making, agriculture, biomass refining, textile, and laundry[34, 36, 37].

1.4.2 Applications of cellulases in various industries

Cellulases have become the third largest group of enzymes used in the industry since their introduction only since a decade[38]. Given below are the applications of cellulases in various industries[39].

Industry	Applications
Agriculture	Generation of plant and fungal protoplasts; plant pathogen and disease control; enhanced plant growth and flowering; enhanced seed germination and improved root system; reduced dependence on mineral fertilizers; improved soil quality
Bioconversion	Conversion of cellulosic materials to bioethanol, organic acids and single cell protein; production of energy-rich and nutritionally improved animal feed; improved feed digestion and absorption; improved ruminant performance; preservation of high quality fodder
Detergents	Cellulase-containing detergents; cleaning action without damaging fibres; improved colour brightness; improved dirt removal; anti-redeposition of ink particles; removal of rough protuberances in cotton fabrics

Food	Antioxidants' release from fruit and vegetable pomace; improved maceration, pressing, and colour extraction of fruits and vegetables; clarification of fruit juices; improvement of yields in starch and protein extraction; improved texture and quality of bakery products; controlled bitterness of citrus fruits; improved texture, flavour, aroma, and volatile properties of fruits and vegetables
Pulp and Paper	Biomechanical pulping; Co-additive in pulp bleaching; pulp freeness and cleanliness; enzymatic deinking; reduced chlorine requirement; reduced energy requirement; improved fibre brightness, strength properties; production of biodegradable cardboard, sanitary paper, and paper towels
Textile	Textile fibre biopolishing; improved absorbance property of fibres; improved fabrics quality and stability of cellulosic fabrics; softening of garments; restoration of colour brightness; removal of excess dye from fabrics

1.4.3 Cellulases: A potential solution to energy crisis

With the advent of industrial revolution and subsequent industrialization hitting the cards, the need for energy was imperative. A popular belief prevailed in Energy fraternity that the conventional petro-fuels would suffice for the global energy demands for a limited period only. With the rapid declination in geological reserves, coupled with environmental concerns, there was an urgent need to discover alternative energy sources. With a view to alleviate the impact of energy crisis, biofuels have been highlighted as an attractive fuel option at the vanguard of alternative energy technologies. These renewable energy sources, together with conventional fuels, are thought to meet the current energy demands. Most of the countries have increased their renewable energy share in their national energy budget in the wake of energy security and climate change and India also has given high priority to renewable energy due to its growing energy demand.

Lignocellulosic biomass (cellulose, hemicellulose and lignin) is the most abundant material on earth. Agricultural residues (e.g. corn stover and wheat straw), forestry residues (e.g. sawdust and mill wastes), portions of municipal solid waste (e.g. waste paper) and various other industrial wastes containing huge

amounts of cellulose have a great potential to be used in the industrial processes. Besides many industrial applications, there is a prospect of converting this biomass into biofuels so that the current energy demands can be fulfilled. But the effective conversion of biomass into biofuels is limited by the nearly impenetrable architecture of plant cell walls, chemical and physical changes to biomass during pre-treatment and structural features of the biomass-degrading enzymes. To make it more difficult, there is no single process which is cost-effective in the degradation of the natural cellulosic biomass and thus the use of the current commercial preparations of cellulase for deconstruction of cellulosic waste is economically not feasible. Cellulases that are more thermostable, halotolerant and that show stability at varied pH can be a major breakthrough in this regard. These properties make them potential candidates for the deconstruction of biomass and production of second-generation biofuels.

Owing to their significance in hydrolysing cellulosic biomass they are extremely valuable for application in plant cell wall deconstruction and generation of second generation biofuels that is critical to the energy supply, environment and economy of the modern world. But the biomass-degrading cellulases studied so far act so slowly that their activity and effectiveness cannot be improved significantly. So there is a need for unlocking novel cellulases from natural resources that could degrade cellulosic biomass effectively and efficiently.

1.5 Rumen microbial ecosystem

Rumen is a complex ecosystem that harbours a wide variety of microorganisms, including bacteria, archaea, protozoa and fungi[40]. A principal function of the microbiota inhabiting rumen is the conversion of plant-based diet into digestible compounds that can be used by the ruminant host[41]. This function is of tremendous importance as it allows the microbes and the host form an interdependent symbiotic relationship in which the host provides the microbes with a suitable habitat for growth, and the microbes, in turn, supply protein, vitamins, and short-chain fatty acids for the animal[42].

The exploitation of rumen ecosystem has been slow, mainly due to the fact that rumen microbial diversity cannot be defined using the traditional culture-based techniques. Enumerating a specific bacterial species and to quantify its role in the rumen fermentation is difficult with the traditional culture-based techniques due to

large number of biochemical tests to be performed and imprecision of the technique even for the most predominant microbe present in the rumen[43]. However, the advent of molecular techniques such as metagenomics has helped immensely in exploring this microbiota-rich resource that harbours a reservoir of industrially important fibrolytic enzymes[44].

1.5.1 Rumen metagenomics and cellulases

Rumen, owing to the presence of unique diverse microbial consortia and microbial products, is one of the best microbial habitats to explore the sources of biomass-degrading enzymes like cellulases that may find use in the biofuel industry[45]. In addition to increasing our knowledge of the biochemical machinery used by the symbiotic animal-microbe consortia for the breakdown of biomass, the metagenomic approach offers new avenues to potentially accomplish biomass conversion at an industrial scale. Just as the knowledge of plant genomes are providing us with new approaches for accelerating feedstock domestication, the metagenomics of biomass-degrading microbes and microbial consortia offer a new means to discover lignocellulolytic enzymes.

Additionally, metagenomics offers a huge potential for discovering novel microbial enzymes with known or novel functions, new proteins with unique functions, and other novel bioactive molecules. Metagenomics, transcriptomics and proteomics of the rumen ecosystem would make the development of functional genomics of important microbes and microbial consortia feasible, which finally aims to characterize the molecular diversity and assessment of various types of interactions, not only between the microorganisms, but with the host as well[46]. Discovering and harnessing the best biomass degrading enzymes directly from natural ecosystems, such as rumen, through molecular techniques ultimately will have a significant impact on increasing the efficiency and reducing the cost of cellulosic biofuel production.

A variety of ruminants have been studied using molecular tools such as domestic cow[47], sheep[48], swamp buffalo[49], and yak[50]. Most studies present in the literature about the rumen microbiome have focused on these ruminants; however, studies applying molecular tools to describe the goat rumen microbial diversity are scarce[51].

1.5.2 Goat rumen

Goats (*Capra hircus*) feed on grasses, the tips of woody shrubs, lignocellulosic agricultural by-products such as cereal straws and stovers. Symbiotic microbes in the rumen of these ruminants play pivotal roles in providing the hosts with various nutrients[52]. Enzymes secreted by ruminal microbes are needed for the conversion of cellulose and hemi-cellulose into simple sugars. These sugars serve as a source of energy for these animals[53]. Thus the goat rumen microbiota is well adapted to convert plant biomass into nutrients. Despite the longstanding interest in the rumen, the literature shows that the goat rumen microbiota is still largely unknown and represents untapped metabolic resources. This reason makes the goat rumen an ideal place to find microorganisms and enzymes specialized in plant cell wall degradation, an area of great interest due to the current efforts to produce second generation ethanol from lignocellulosic feedstocks[54].

Goat rumen contains partially digested cellulosic biomass which is expected to harbour bacteria producing cellulase enzyme. Accordingly, the goat rumen digesta is used for the assessment of gene(s) capable of encoding the desired cellulase enzymes. The gene in turn can be cloned in culturable bacteria like *E. coli* BL21(DE3) for the production of the enzyme.

1.6 Objectives of the study

On the basis of the background information, the present research work was taken up for exploiting the metagenomic approach to search for the industrially important cellulase(s) with the following objectives:

1. Isolation and purification of metagenomic DNA from environmental sample followed by the identification of the gene for desired enzyme using metagenomic approach.

2. Cloning the gene of interest into a suitable expression vector and transformation of desired enzyme producing gene to *E.coli*.

3. Identification of positive clones expressing the gene of interest using function-driven approach of metagenomics and assessment of production of desired enzyme.