Chapter 2: Review of literature

2.1 Microbial diversity

Microbes are the most numerous, diverse and dynamic organisms constituting the major reservoir of genetic diversity on earth[1, 55]. They are ubiquitous in every habitat such as soil[56], water[57], air[58], gut of humans[59] and animals[60]; inhabit highly polluted environments like the petroleum-contaminated soils[61], heavy-metal-contaminated sewage sludge[62]; tolerate most extreme conditions such as acidic hot springs[63], Antarctic environments[64], deep sea hydrothermal vents[65], rocks in deep boreholes beneath the earth's surface[66] etc. In fact they are everywhere, occupying every part of the earth and managing its geochemistry, cycling of elements and breakdown of wastes.

According to Whitman *et al.*[1], the total number of bacterial and archaeal cells on earth has been estimated to be 4-6 x 10^{30}. It is interesting to know that we, humans, have a greater number of bacterial cells (10^{14}) than human cells (10^{13}) in our bodies[67]. These millions and billions of benign microbes that live in our gut help us to digest food, break down toxins, and fight off pathogenic microbes. They also clean up pollutants in the environment, such as oil and chemical spills. All these activities are carried out by complex microbial communities—intricate, balanced, and integrated entities that adapt swiftly and flexibly to environmental changes[68].

2.1.1 The uncultivable majority

In 1980's when the field of microbiology was booming, Staley and Konopka[69] came up with the observation that there lies a huge difference between estimated population sizes when using direct microscopic cell counts and the number of colonies on nutrient agar. This "great plate count anomaly" along with other studies[70, 71] substantiated the claim that a vast majority of bacteria are uncultured.

There might probably be a variety of reasons as to why a major fraction of microorganisms are recalcitrant to laboratory cultivation. Most of these can be reduced to the difficulty of replicating the very precise environmental conditions

certain microbes require for growth; and for others the interdependence with other organisms in the natural ecosystem might be crucial for their existence.

Cho and Giovannoni[72] reported many attempts that have recently been made to cultivate previously uncultured microbes using novel approaches. These include high-throughput culturing (HTC) using dilution-to-extinction[73], cultivation with a diffusion growth chamber[74], encapsulation of cells in gel micro-droplets[75], and modified plating methods[76, 77]. However, only about 0.01 to 0.1% of all the microbial cells from all environments tend to form colonies on standard agar plates; thus a major fraction still remains uncultivated, suggesting that further innovation will be needed for their successful cultivation. To explore the huge untapped resources of genetic diversity, other methods are therefore important.

In 1985, a culture-independent method was introduced by Lane *et al.*[78] that involved direct analysis of 16S rRNA gene sequences to describe the diversity of the microorganisms present in environmental samples. These, together with the development of the Polymerase Chain Reaction[79], were advances that radically changed our understanding of the microbial world. Handelsman[3] reported high degree of divergence of the uncultured bacteria from that of the cultured bacteria on the basis of direct DNA analysis of environmental samples with sequencing.

2.2 Metagenomics

To explore the hidden and untapped genetic resources, Handelsman *et al.*[80] introduced the concept of 'metagenomics', a culture-independent approach which involves genomic analysis of DNA extracted directly from its natural environment. This is a more inclusive strategy to access the total microbial genetic reservoirs as compared to the traditional, culture-dependent approaches. Metagenomics involves genome library construction, followed by sequencing and screening it for novel pathways and biomolecules. It differs from traditional genomic library construction in that the cloned DNA does not originate from a single known microbe, but rather from the entire population in an environmental sample. Metagenomics allows the assessment and characterization of the taxonomic as well as the metabolic diversity of microbial communities in a highly extended fashion as compared to the traditional culture-based methods.

2.2.1 Rumen metagenomics

The exploitation of rumen ecosystem has been slow, mainly due to the fact that the rumen microbial diversity cannot be defined using the traditional culture-based techniques. 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]. Rumen metagenomics involves both function-based and sequence-based approaches.

2.2.1.1 The function-driven approach

The function-driven approach of metagenomics is based on the identification of the clones that express a desired trait or phenotype such as antibiotic resistance, growth on an unusual substrate, or enzyme production, followed by biochemical and molecular characterization of the active clones. This is a quick approach that identifies the clones expressing bioactive compounds, enzymes or other molecules that have industrial, medicinal or agricultural applications. Gastrointestinal (GI) tract has been one of the most sought-after habitats to explore the hitherto uncultured microbial diversity for the novel enzymes and bioactive compounds. Lan *et al*.[81] isolated novel bacterial strains from the caecum of chicken and characterized them by determining their phenotypic characteristics, cellular fatty acid profiles, menaquinone profiles and phylogenetic positions based on 16S rRNA gene sequence analysis. Ferrer *et al*.[82] reported biochemical and structural features of a novel cyclodextrinase from cow rumen metagenome using functional metagenomic approach. Feng *et al*.[83] isolated eleven independent clones expressing cellulase activities from rabbit cecum by using functional metagenomics. The encoded products shared less than 50% identities and 70% similarities to cellulases in the databases. Beloqui *et al*.[84] used function-based approach of metagenomics to mine a novel polyphenol oxidase from a metagenome expression library of bovine rumen.

2.2.1.2 The sequence-driven approach

Sequence-based metagenomics may be used either to study the microbial community in a particular habitat, or to search for genes that encode novel protein or bioactive molecules. Sequence conservation of regions of phylogenetic anchors such as 16S ribosomal RNA (rRNA), recA, rpoB etc. facilitates their isolation

without prior knowledge of full gene sequence. Thus, application of this approach is likely to be broadened as patterns emerge that define more gene families containing highly conserved features.

Detailed information of the microbial community composition in natural ecosystems can be gained from the phylogenetic analysis of 16S and 18S rRNA gene sequences obtained directly from environmental samples by PCR amplification, cloning and sequencing. An important application of rumen metagenomics is the identification of members of rumen microbiota which have not been cultured till date. Leser *et al*.[85] reported an inventory of phylotypes in the GI tracts of a collection of Danish pigs. Their results documented a hitherto unknown bacterial diversity indicating that the majority of intestinal bacteria are uncharacterized. Ferrer *et al*.[86] constructed a metagenomic DNA library extracted from the rumen contents of dairy cattle. Sequence analysis revealed that 36% (8/22) of the enzyme-encoding gene sequences were entirely new and formed deep branched phylogenetic lineages with no close relatives among known esters and glycosyl hydrolases. Nicholson *et al*.[87] reported a temporal temperature gradient gel electrophoresis method to determine the methanogen diversity in cattle and sheep rumen showing that uncultured methanogens had significant population densities in each of the rumen samples examined. In a breakthrough study, Edwards *et al*.[88] reported that 89% of the microbial diversity present in the rumen had greatest similarity to organisms which had not been cultured, and that several sequences were likely to represent novel taxonomic groupings. This meant that only about 11% of the microbial diversity was represented by cultured isolates (based on >95% 16S rDNA identity). The study therefore emphasised on the need to reconcile traditional culture-based rumen microbiology with molecular ecological studies to determine the metabolic role of hitherto uncultivated microbial species.

2.3 Goat rumen metagenomics

Goat (*Capra hircus*) is a member of a class of animals called ruminants who chew their cud (ruminate). Unlike humans, they have a special four-compartment stomach, consisting of the rumen, the honey-combed reticulum, omasum, and the abomasum or true stomach. This four-chambered stomach is

especially designed to digest roughage such as grass, hay and silage. Goats feed on the tips of woody shrubs, cereal straws and stovers. Symbiotic microorganisms inhabiting the rumen play pivotal role in providing the hosts with various nutrients[52]. Miyagi *et al*.[53] reported that the 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 ruminants. There are also fermentative microbial populations that transform simple sugars into low molecular weight fatty acids, which are also used as energy source by the ruminants. Thus the goat rumen microbiota is well adapted to convert plant biomass into nutrients. This reason makes the goat rumen an ideal place to find microorganisms and enzymes specialized in functions as diverse as plant cell wall degradation, an area of great interest due to the current efforts to produce second generation ethanol from lignocellulosic feedstocks, xylanase and protease production etc[54].

In view of the above observations, many studies have focussed on exploring the untapped microbial diversity present in the goat rumen. Cheng *et al*.[51] reported the molecular diversity analysis of rumen methanogenic Archaea from goat by DGGE methods using different primer pairs. From a set of primer pairs used, they concluded that the procedure of DGGE analysis with primer pair 519f/915rGC was more suitable for investigating methanogenic archaeal diversity in the rumen. The dominant methanogenic Archaea in the goat rumen was *Methanobrevibacter* sp. and an unidentified methanogenic *Archaea*.

Lim *et al*.[89] carried out the metagenomic analysis of microbial community structure and specific protein domains with cellulase activity in the goat rumen using shotgun-sequencing and bioinformatics. The presence of specific dominant bacterial species such as *Prevotella ruminicola*, *Butyrivibrio proteoclasticus*, and *Butyrivibrio fibrisolvens* was observed among 1,431 bacteria in the goat rumen fluid. In addition to that, 28 protein domains with cellulase-like activity such as lipase GDSL, cellulase, and Glycosyl hydrolase family 10 proteins were identified with strong positive correlation which was indicative of microbial adaption in the goat rumen based on feeding habits.

Cunha *et al*.[54] reported the characterization of bacterial and archaeal communities present in the liquid- and solid-associated fractions of the rumen from

Moxotó breed goats using 16S rDNA sequencing. They found that the overall dominant classes in the rumen were *Clostridia* and *Bacteroidia*, which are known to play an important role in plant fibre degradation in other ruminants. 4.7% and 16.4% of the liquid-fraction and solid-fraction sequences, respectively, belonged to unclassified bacteria. From the archaeal libraries they could identify only the sequences from the phylum *Euryarcheota* and assigned them to the class *Methanobacteria* belonging to the genera *Methanobrevibacter* and *Methanosphaera*. The presence of these groups in the Moxotó goat rumen is justified because the local water contains high salt concentrations.

In addition to determining the microbial diversity of goat rumen, several metagenomics-based studies have focussed on exploiting the goat rumen contents for novel enzymes that may find use in biotechnological and industrial fields. Pushpam *et al*.[90] isolated and characterized an alkaline serine protease from goat skin surface metagenome which showed homology to peptidase S8 and S53 subtilisin kexin and sedolisin of *Shewanella* sp. Surprisingly, this alkaline serine protease requires Co^{2+} or Mn^{2+} metal ions for its improved activity that makes it a suitable candidate for the potential applications in the detergent and laundry industries.

Wang *et al*.[91], for the first time, reported cloning and characterization of a novel arylesterase from the goat rumen contents. Biochemical characterization of the enzyme showed that it was an arylesterase with broad substrate specificity, high catalytic efficiency, stability at mesophilic temperatures, and strong resistance against protein-denaturing compounds suggesting that the enzyme could find potential applications in the biotechnological and industrial fields.

Again, Wang *et al*.[92] reported a novel xylanase gene encoding Glycosyl Hydrolase family 10 (GH 10) xylanase cloned directly from the goat rumen. Biochemical characterization and structure analysis showed it to be a cold active xylanase having a higher catalytic efficiency and displaying better thermostability at mesophilic temperatures than other known GH 10 cold active xylanases. These properties make the enzyme an alternative to feed and food additives currently being used.

Cheng *et al*.[93] reported the isolation and characterization of a non-specific endoglucanase from a goat rumen metagenomic library. The amino acid sequence

of the enzyme was homologous with cellulases belonging to the glycosyl hydrolase family 5. The expressed protein showed activity towards carboxymethyl cellulose (CMC) and xylan, suggesting non-specific endoglucanase activity and thus could be used as a potential candidate for feed additive.

2.4 Bioinformatics tools for metagenomic data annotation

The interpretation of metagenomic-derived clone sequences and their subsequent characterization and taxonomic classification demands new computer programs and tools that ease the annotation of metagenomic data. Huson *et al.*[94] developed a new computer program MEGAN (MetaGenome ANalyzers) that generates specific profiles from sequencing data after assigning the reads to NCBI taxonomy using a straight-forward algorithm. The MEGAN approach has been found applicable to several data sets including subset of Sargasso Sea data set (obtained by Sanger sequencing), data obtained from mammoth (*Mammuthus primigenius*) bone (obtained by "sequencing-by-synthesis" approach) and identifying the microbial species based on already available microbial (*Escherichia coli* and *Bdellovibrio bacteriovorus*) genome sequence information[44].

Meyer *et al.*[95] reported MG-RAST (Rapid Annotation using Subsystems Technology for Metagenomes) a high-throughput pipeline developed to provide high-performance computing to researchers using metagenomics. The server provides several methods to access the different data types, including phylogenetic and metabolic reconstructions, and the ability to compare the metabolism and annotations of one or more metagenomes and genomes.

Caporaso *et al.*[96] described 'quantitative insights into microbial ecology' (QIIME), an open-source software pipeline to address the problem of taking sequencing data from raw sequences to interpretation and database deposition. QIIME (http://qiime.sourceforge.net/) supports a wide range of microbial community analyses and visualizations that have been central to several recent high-profile studies, including network analysis, histograms of within- or between-sample diversity and analysis of whether 'core' sets of organisms are consistently represented in certain habitats. QIIME also provides graphical displays that allow users to interact with the data. It allows alternative components for functionalities such as choosing operational taxonomic units (OTUs), sequence alignment,

inferring phylogenetic trees and phylogenetic and taxon-based analysis of diversity within and between samples to be easily integrated and benchmarked against one another.

Albanese *et al*.[97] introduced MICCA, a software for taxonomic profiling of metagenomic data, for the processing of amplicon metagenomic datasets that efficiently combine quality filtering, clustering of OTUs, taxonomy assignment and phylogenetic tree inference.

The tools and approaches developed so far are applicable for any metagenome project, irrespective of sequence type, thus removing a primary bottleneck in metagenome sequence analysis – the availability of high-performance computing for annotating the data.

2.5 Metagenomics as a tool for the discovery of industrially important enzymes

The construction and screening of metagenomic libraries has turned out to be a powerful tool in exploring untapped microbial diversity for novel enzymes such as cellulases, lipases, amylases, proteases etc. and secondary metabolites such as antibiotics, as it enables the complete microbial diversity to be investigated and not only the culturable proportion of about 0.1 to 1%[98]. Several metagenomic studies have resulted in finding biocatalysts with interesting features (Table 1).

Table 2.1 Biocatalysts screened from metagenomic libraries (Bashir *et al*.[8])

Function	Habitat	Library type	Substrate	Reference
Alkaline protease	Goat skin	Plasmid	Skim milk agar	Pushpam *et al*.[90]
Amylase	Soil	Cosmid	Starch	Sharma *et al*.[99]
α-amylase	Soil	Fosmid	Starch	Vidya *et al*.[100]
Cellulase	Pig manure	Fosmid	Carboxymethyl-cellulose	Kwon *et al*.[101]
Cellulases	Biogas plant	Fosmid	Carboxymethyl-cellulose	Ilmberger *et al*.[102]

Cellulase	Biogas digester	Fosmid	Carboxymethyl-cellulose and β-D-glucan	Geng *et al.*[103]
Cellulase	Rabbit cecum	Cosmid	Carboxymethyl-cellulose, Esculin hydrate	Feng *et al.*[83]
Cellulase	Soil	Fosmid	Hydroxyethyl cellulose	Nacke *et al.*[104]
Endocellulase	Soil	Plasmid	Carboxymethyl-cellulose	Archana *et al.*[105]
Endoglucanase	Compost soils	Cosmid	Carboxymethyl-cellulose	Pang *et al.*[106]
Endoglucanase	Rice straw compost	λ phage	Carboxymethyl-cellulose	Yeh *et al.*[107]
Esterase	Soil	Fosmid	Tributyrin	Zhang *et al.*[108]
2, 4-D hydroxylase	Soil	Plasmid	3,5-dichlorocatechol	Lu *et al.*[109]
Lipase	Lagoon	Fosmid	Triolein	Glogauer *et al.*[110]
Metallo-protease	Compost soil	Plasmid	Skim milk agar	Waschkowitz *et al.*[111]
Pectinase	Soil	Plasmid	Polygalacturonic acid	Singh *et al.*[112]
Serine proteases	Sand	Plasmid	Skim milk agar	Neveu *et al.*[113]
Xylanase	Soil	Plasmid	Xylan	Hu *et al.*[37]
Xylanase	Pig manure compost	Fosmid	Xylan	Kwon *et al.*[101]

2.6 Biotechnological relevance of cellulases

Due to the increasing oil prices, environmental problems and uncontrolled exhaustion of fossil fuels the demand for energy sources alternative to fossil fuels

is rising. Lignocellulosic biomass has been projected as a potential source for the production of second-generation biofuels, an alternative renewable energy that may meet the rising energy demand. Bioethanol, a second-generation biofuel can be generated from cellulosic material like wood or grass or from crop waste and forestry residues. For that, the lignocellulosic substrate must first be hydrolysed to glucose which can be fermented to ethanol. The hydrolysis of cellulose, the most abundant biopolymer on earth, can be performed either chemically[114] or enzymatically[115]. The chemical process includes a combination of heat, pressure and acids. The major disadvantages of chemical hydrolysis of cellulose include the high energy input and the formation of toxic waste products which have to be depolluted cost-intensively. The main drawback of the enzyme-based hydrolysis is that cellulose is insoluble in water and that the cellulolytic enzymes need an aqueous medium for the biotechnologically relevant conversion rate. Thus, cellulases with better properties such as tolerance to ionic liquids, high thermostability, wide range of pH tolerance and efficient conversion rate are desired.

In addition to the production of biofuels, cellulases find biotechnological applications in giving jeans a "used-look". They are also used in the pulping of fruits resulting in an improved extraction of juice[116] and as additives in laundry agents[117]. Cellulases are also used for the de-inking of recycled fibres in paper industry and for biomechanical pulping[116].

2.7 Metagenomics for cellulases

Due to the expansion of the cellulosic biofuel industry and an ever-increasing demand for alternative energy, interest in cellulolytic enzymes has increased in recent years. The rumen, a highly adapted environment for the efficient degradation of cellulose, is considered to be a promising source of enzymes for industrial use. Several studies have been undertaken for functional metagenomic screening so as to identify cellulase enzymes that might be of such use from the microbial community present in the environmental samples. Pang *et al.*[106] prepared a cosmid metagenomic library using DNA isolated from the compost soils. Functional screening of the library resulted in four novel cellulase genes; one of these genes, umcel9B was expressed in *E. coli* M15 (pREP4) and the purified recombinant enzyme was found to be active at low temperatures

exhibiting the properties similar to the cold-active endoglucanase. Cold-active enzymes are attractive for their values in biotechnological applications.

Liu *et al.*[24] reported a novel endo-β-1,4-glucanase from a soil metagenomic library. It exhibited high activity at low temperature, pH and thermal stability, halo-tolerance, stability in the presence of proteolytic enzymes, and high V_{max}. These characteristics suggested the novel metagenome-derived cellulase to be a potential candidate for the industrial applications.

Alvarez *et al.*[118] described the discovery and characterization of a novel member of GH16 family from the sugarcane soil metagenome. The enzyme contained 286 amino acid residues and displayed sequence homology and activity properties that resemble known laminarases. According to them SCLam was the first nonspecific (1,3/1,3:1,4)-β-D-glucan endohydrolase recovered by metagenomic approach to be characterized.

The rumen ecosystem has also been exploited for discovering the biomass degrading cellulases. Duan *et al.*[119] identified novel cellulases with distinct properties from metagenomes of buffalo rumens. They provided evidence for the diversity and function of cellulase and mechanisms of cellulose hydrolysis in the rumen. According to the study, thirteen recombinant cellulases were partially characterized showing diverse optimal pH from 4 to 7. Seven cellulases were found to be most active under acidic conditions with optimal pH of 5·5 or lower. Cellulases cloned in their work probably played important roles in the degradation of celluloses in the variable and low pH environment in buffalo rumen.

Gong *et al.*[120] reported a cellulase gene, Cel14b22, which was expressed at a high level in *E. coli*. The purified enzyme was found to be stable over a broad pH range (from 4.0-10.0) and its activity was significantly enhanced by the presence of Mn^{2+} and reduced by the presence of Fe^{3+} or Cu^{2+}. The enzyme hydrolysed a wide range of β-1,3-, and β-1,4-linked polysaccharides. Microcrystalline cellulose and filter paper were efficiently degraded by the metagenomic-derived recombinant enzyme.

Loaces *et al.*[121] cloned and expressed in *E. coli* a metagenomic-derived cellulolytic enzyme from cow rumen with both glucanase and xylanase capabilities. It was found to be related to a yet non-sequenced microorganism related to *Prevotella*. This novel cellulase was active at low temperatures and

exhibited thermostability, tolerance to high concentrations of inorganic salts, imidazolium ionic liquids (ILs), inhibitors such as furfural or acetate and final products viz. glucose and cellobiose.

Li *et al*.[122] reported the construction of a metagenomic fosmid library using genomic DNA isolated from the ruminal contents of four adult gayals, a rare semi-wild bovine species found in Indo-China. This library contained 38,400 clones with an average insert size of 35.5 kb. Investigation of the library for cellulase activity led to the identification of the Umcel-1 gene. The putative Umcel-1 gene product belonged to the glycosyl hydrolase family 5 and showed the highest homology to the cellulase from *Clostridium lentocellum*. The purified recombinant Umcel-1 hydrolyzed carboxymethyl cellulose with optimal activity at pH 5.5 and 45°C. According to them, their study provided the first evidence for a cellulase produced by bacteria in gayal rumen.

The goat rumen microbiota still largely remains to be unknown. The goat rumen is 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.