2.1 Microbial diversity

Biodiversity of India is a unique asset, which can be harnessed through use of modern biotechnological tools of research. Microorganisms have been evolving for nearly 4 billion years and are capable of exploiting a vast range of energy sources and thriving in almost every habitat like surviving and thriving in extremes of heat, cold, radiation, pressure, salt, acidity and darkness. In these environments no other forms of life are found and the only nutrients come from inorganic matter. For 2 billion years microbes were the only form of life on Earth. During this long history, all of the basic biochemistries of life evolved, and all life forms have developed from these microbial ancestors. It is estimated that 50% of the living protoplasm on this planet is microbial (Whitman et al., 1998). Microorganisms represent by far the richest repertoire of molecular and chemical diversity in nature. They underlie basic ecosystem processes such as biogeochemical cycles and food chains, as well as maintain vital and often elegant relationships between themselves and higher organisms. Microbes provide the fundamental underpinning of all ecosystems.

Diverse microorganisms are essential to sustainable biosphere. They are able to recycle nutrients, produce and consume gases that affect global climate, destroy pollutants, treat our wastes and they can be used for biological control of plant and animal pests. A large population of microbes can be rapidly scanned for their taxonomic grouping revealing new microbes, which have not been separated till date. These can be checked for their special characteristics like secondary metabolites for e.g., antibiotics, hydrogen, methane, bioplastic, and bioremediation of oily sludge, biowastes. These and other properties can prove beneficial for improving the environment and consequently human health (Kalia et al., 1994; 2003 a, b; 2007; Kalia and Purohit, 2008). Without microorganisms, all life on earth would cease. Microbial ecology studies reveals diversity of microorganisms and how micro-organisms interact with each other and with their environment to generate and to maintain such diversities (Konstantinidis et al., 2006).

Consequently, microbiology has focused on two areas of study: (i) microbial diversity, including the isolation, identification and quantification of micro-organisms in various habitats; and (ii) microbial activity, that is, what micro-organisms are doing in their habitats and how their activities contribute to the observed microbial diversity. The
patterns of distributions are often studied in the context of environmental factors such as
temperature, pH, salinity, pressure, the availabilities of water and nutrients, and the
sources of energy like carbon and nitrogen. These ecological factors influence microbial
activities and play very important roles in determining the dynamics of micro-organisms
in natural environments (Fenchel, 2005).

The microbial diversity represents metabolic, structural, morphological and
 genetic diversity. Conventional methods for isolating and identifying microbes have been
 a limiting factor for exploring and exploiting microbial wealth. These methods have not
 been able to deliver new enzymes, molecules, products, antibiotics etc. due to very
 limited variability. Since, we are not able to culture majority of these organisms, there is
 a demand for novel techniques and effective strategies to tap the unexplored diversity.

Also the existing methods for culturing microorganisms limit analysis to those
 that grow under laboratory conditions. The recent surge of research in molecular
 microbial ecology provides compelling evidence for the existence of many novel types of
 microorganisms in the environment in numbers and varieties that dwarf those of the
 comparatively few microorganisms amenable to laboratory cultivation. Corroboration
 comes from estimates of DNA complexity and the discovery of many unique 16S rRNA
gene sequences from numerous environmental sources. Collectively, the genomes of the
total microbiota found in nature, termed as metagenome (Handelsman et al., 1998)
contain vastly more genetic information than is contained in the culturable subset. Given
the profound utility and importance of microorganisms to all biological systems, methods
are needed to access the wealth of information within the metagenome. The information
gained from the metagenomic procedure provides information regarding the structure,
organization, evolution and origin of the DNA and can be used in scientific applications
for the benefit of society and the environment as industry, therapeutics, and
environmental sustainability.

Biological and environmental research supports innovative high-impact and peer-
reviewed science. To aid in carrying out its missions, scientists expect to find a vast
reertoire of useful functions in microbial world that could be applied to solving
challenges in human world. Biological, environmental research and closely linked
Genomics program are generating novel insights into both biological underpinnings of
climate change and the role of microbes in the overall processing of metals, carbon, radio nuclides and nitrogen (Ellis et al., 2003). People are only beginning to appreciate the power of microbial gene sequencing for advancing science. Sequencing of most of interesting microbes is done to uncover their potential applications.

2.2 Phenotypic methods

The earliest methods that were used to identify and type organisms were based upon their phenotypic characteristics.

2.2.1 Biotyping

One of the most widely utilized techniques is biotyping, or the differentiation of strains based on properties such as differences in biochemical reactions, morphology, and environmental tolerances. Biotyping is often used to help determine the species of microorganisms based upon their abilities to utilize components in different growth media and carry out certain chemical reactions, but it can also be used to separate members of a particular species due to biochemical differences among the organisms. Biotyping is now routinely performed in laboratories using automated systems designed for species identification (Singh et al., 2006).

2.2.2 Antimicrobial susceptibility

This testing is a common practice in the clinical microbiology laboratory. The resultant antibiogram indicates the pattern of in vitro resistance or susceptibility of an organism to a panel of antimicrobial agents. According to Pankuch et al., (2006) agar dilution, micro-dilution, E-test and disk diffusion are satisfactory methods for susceptibility testing. Disk diffusion methods are not used as commonly as they once were because of the lack of automation for testing. Microdilution testing provides a quantitative measure of the MIC, which is defined as the lowest concentration of the antimicrobial agent that inhibits the growth of the organism. Both disk diffusion and broth dilution have been carefully standardized and are therefore quite reproducible within and between laboratories. In most epidemiologic studies the antibiogram
has limited value because isolates that are not genetically and epidemiologically related may have the same susceptibility pattern.

### 2.2.3 Serotyping

It uses a series of antibodies to detect antigens on the surface of bacteria that have been shown to demonstrate antigenic variability (Babl et al., 2001). Serotyping methods have been used for decades for the taxonomic grouping of a number of bacterial pathogen species and remain important for typing *Salmonella*, *Legionella*, *Shigella*, and *Streptococcus pneumoniae* isolates. Serotyping also has been shown to have epidemiologic value in differentiating strains within species of nosocomial pathogens such as *Klebsiella* and *Pseudomonas*. There are a number of different ways in which serotyping can be performed; each varies the way in which the antibody-antigen reactions are detected. It is known that serotype distribution of strains causing invasive disease, nasopharyngeal colonization and antibiotic resistance are linked to age, geography, and socio-economic conditions of that population (Aslan et al., 2007). Analysis of the leading serotypes in a particular area is important to evaluate the efficacy of vaccines.

### 2.2.4 Phage and bacteriocin typing

Bacteriophage and bacteriocin typing as epidemiologic tools are limited to bacteria. Bacteriophage (phage) typing classifies bacteria based on the pattern of resistance or susceptibility to a certain set of phages (Hopkins et al., 2004). Bacteriophages are viruses that are able to attach to the cell walls of certain bacteria, enter, multiply, and lyse the cells. The differential ability of phages to infect certain cells is based upon the availability of corresponding receptors on the cell surface for the phage to bind. Bacteriophage typing has some drawbacks due to a lack of widespread availability of biologically active phages and the technical difficulty of performing the technique, but the method has been applied to a number of bacteria associated with nosocomial infections. Additionally strains can be typed based on their susceptibility to a set of heterogeneous substances (generally proteins) that are produced by other bacteria. These inhibitory compounds, or bacteriocins, often limit the growth of closely related
2.2.5 Limitations of phenotyping methods

The use of phenotypic methods for the characterization of nosocomial pathogens has been useful for our understanding of pathogens; however, these methods have drawbacks that limit their utility for highly discriminatory typing of microorganisms. Limitations of serotyping include a lack of availability of certain antisera and problems with standardization of different methods. Biotyping often lacks discriminatory power because of variations in gene expression and random mutations that may alter biological properties of microorganisms. Biotyping cannot differentiate among strains where biochemical diversity is uncommon, such as the enterococci, and therefore the utility of biotyping in epidemiologic studies is quite limited. Bacteriophage typing is labor-intensive, and the method often demonstrates poor reproducibility and standardization. When other phenotyping methods fail, bacteriocin typing may have some utility for organisms not easily typed, such as \textit{P. aeruginosa} and \textit{Candida} species. Despite these limitations, phenotypic characterization continues to play a vital role in the overall management of infectious diseases (Singh et al., 2006).

2.2.6 Why molecular methods

Because of the inherent limitations of conventional phenotyping methods for detecting circulating strains, as well as their mechanisms of resistance, molecular techniques that complement the information provided by these methods have been developed (Pérez-Trallero et al., 2007). The application of molecular biological methods to study the diversity and ecology of microorganisms in natural environments has been practiced since the mid-1980s. Since that time many new insights into the composition of uncultivated microbial communities have been gained. Whole groups of organisms that are only known from molecular sequences are now believed to be quantitatively significant in many environments. Molecular methods vary with respect to discriminatory power, reproducibility, ease of use, and ease of interpretation (Lasker, 2002; Hall, 2007). Molecular methods have also allowed characterization of many long-recognized but
poorly understood organisms (Head et al., 1998). Developments in microbial ecology and industrial biotechnology are severely hampered by the lack of a reliable identification system. It is notoriously difficult and often unsatisfactory to try to identify fresh bacterial isolates from natural environments such as soils, lakes, and oceans. To obtain a conventional phenotypic description requires long and fastidious work, and identifications of novel isolates by the use of dichotomous keys or computerized numerical taxonomy do not warrant satisfactory identifications of species already described in reference manuals or appropriate propositions for actual new bacterial species and genera. On the other hand, phylogenetic relationships among various organisms, and thus their identification, can be derived from the degree of relatedness of their genomes. Methods that use this approach comprise measurements of DNA relatedness over the entire genome, comparisons of restriction patterns, especially ribotyping, and comparative analyses of sequences of homologous genes. DNA-DNA relatedness and ribotyping are best suited for the identification of closely related species or strains within a single species. Presently, a direct comparison of rRNA sequences is probably the most powerful tool for the identification of many bacteria. Indeed, rRNA genes (rDNA) are present in all bacterial species, are truly homologous in all organisms, are easily sequenced, and now offer a large and ever increasing database of sequences and allow the identification of uncultured bacteria.

Studies of microbial isolation, identification, and characterization have always been intimately entwined. Comparative analysis of rRNA sequences not only provided the phylogenetic framework which was lacking in microbial diversity, but also allowed the development of tools to address this vast microbial wealth. The ubiquity of rRNA molecules (small subunit, 16S, large subunit 23S) in all cellular life forms and comparative analysis of their sequences can be universally applied to infer relationships among organisms. The rRNA molecules comprise highly conserved sequence domains interspersed with more variable regions. Essential rRNA domains are conserved across all phylogenetic domains, thus “universal” tracts of sequences can be identified. The most commonly used form of comparative rRNA sequence analysis involves the construction of phylogenetic trees. Ribosomal RNA sequence analyses have been greatly facilitated by the availability of an excellent, indispensable, curated database of rRNA
sequences (the ribosomal database project, RDP-II) (Maidak et al., 2000; Cole et al., 2006).

The development of robust and simple DNA cloning techniques and PCR, have allowed higher resolution analyses of more complex communities using SSU rRNA sequence analysis. The presence of universally conserved sequences at the 5’ and 3’ ends allows the recovery of rRNA sequences and amplification of nearly complete SSU rRNA genes from DNA extracted from natural samples. This is currently the most widely adopted method of sequence retrieval from natural samples. The extracted DNA is subjected to PCR amplification using “universal” primers or primers designed to amplify rRNA genes from a particular group of organisms. The PCR product can be cloned by using commercially available kits for the cloning of PCR products (Insta TA cloning kit). Screening clone libraries for rRNA genes, once cloned, the 16S rRNA gene library can be screened by a variety of methods. Colony hybridization procedures using rRNA gene specific oligonucleotide probes of defined phylogenetic resolution may be used. Plasmid minipreps and restriction digests can be used to confirm the presence of cloned DNA of the correct size, or colony PCR (using sequencing primers with priming sites that flank the insert DNA) can be used as a rapid screening procedure to detect cloned PCR products and can also rapidly provide template DNA suitable for sequencing of specific clones.

Initial screening of rRNA gene containing clones, of purified plasmid DNA or insert DNA obtained by PCR for the presence of near identical sequences, can greatly reduce the number of clones that require complete sequencing. The rapidly expanding database of rRNA sequences, makes it possible to design oligonucleotides of varying phylogenetic resolution. These can be utilized in the detection and enumeration of specific groups of bacteria in clinical specimens, foodstuffs and environmental samples. Statistical methods based on SSU rRNA sequence estimations have been also used (Degrange and Bardin, 1995), in an attempt to obtain quantitative data from PCR-based analyses. There is growing interest in research and development to develop novel tools to study, detect, and characterize microbes and their communities in industrial environments (Maukonen et al., 2003). Molecular approaches can now be considered an essential component of the complete research microbiologist’s toolbox. The widespread use of
molecular approaches is here to stay, and these technologies are at the cutting edge of science.

This work will, however, concentrate on the approaches involving culture dependent, culture independent and phylogenetic analysis of ribosomal RNA sequences.

### 2.3 Genotyping techniques

In recent years, molecular or genotypic techniques have received increased attention as means of analyzing interrelationships. Some of the methods are discussed below.

#### 2.3.1 Polymerase chain reaction (PCR)

PCR is a biochemical in vitro reaction that permits the synthesis of large quantities of a targeted nucleic acid sequence. The procedure requires template DNA from the organism being typed, two complementary oligonucleotide primers that are designed to flank the sequence on the template DNA to be amplified, and a heat-stable DNA polymerase. The PCR primers serve as the starting point for the polymerase to add the bases that make up a strand that is complementary to the template. A growing number of organisms have been studied using this approach (Grimm et al., 2004; Hussain et al., 2004; Suzuki et al., 2004; Trad et al., 2004; Nakari et al., 2005; Shutt et al., 2005). Each amplification cycle consists of a heat denaturation phase in which double-stranded DNA is melted into single strands, an annealing phase where the primers bind to the single-stranded target sequences, and an extension phase. It is during the extension phase, in which the copy number of the DNA is doubled, that the DNA synthesis proceeds from the primers along the template strands, generating copies of the original double-stranded DNA molecule.

#### 2.3.2 Multiplex PCR

In order to increase the efficiency of PCR typing and reduce reagent costs, multiple sets of primers can be included in a single reaction tube in a process termed multiplex PCR (Francois et al., 2004; Focuault et al., 2005). A key strategy in the development of a multiplex PCR assay is the design of the primers. Primers must be
designed such that all of the primers have very close annealing temperature optimums, and the amplification products that they produce need to be of noticeably different sizes to facilitate interpretation. If the amplification products were too close in size, it would be difficult to determine the identity of the amplification product. An additional concern with multiplex PCR is that the mixing of different primers can potentially cause interference in the amplification process, thus making optimization of the reaction difficult, especially as the number of primer pairs in the reaction mixture increases. The development of multiple assays such as multiplex PCR means that several bacterial species can be identified in a single assay (Nocker et al., 2007).

2.3.3 Nested PCR

When there is an extreme need for sensitivity and specificity in PCR, the process of nested PCR can be carried out. Nested PCR involves the sequential use of two PCR primer sets. The first primer set is used to amplify a target sequence (which increases the sensitivity for the second primer set); the amplicon generated then serves as the template for a second amplification using primers internal to those of the first amplicon. This secondary amplification proceeds only if the intended target was initially amplified; if the primary amplification was nonspecific, the secondary amplification would not occur (increased specificity) (Singh et al., 2006). A major drawback of nested PCR is that the reaction vessel needs to be opened in order to add the second primer set, increasing the potential for contamination of the work environment with amplified DNA.

2.3.4 Real-time PCR

The advent of molecular methods such as real time polymerase chain reaction has allowed improvement of detection methods currently used in laboratories (Trani et al., 2006). The amount of PCR product is measured at each cycle and also during the exponential phase, which enables the quantification of the initial template amount. The realtime measurement is based on fluorescent dyes that either bind to double-strand DNA or hybridize to a specific sequence. Since real-time PCR is especially vulnerable to inhibitory compounds, internal standards should always be used when complex sample matrixes are studied.
2.3.5 Hybridization

Hybridization techniques can be used in bacterial identification either alone or combined with a preceding PCR step. In hybridization, a labeled probe (a denatured DNA fragment varying in size between tens of basepairs to kilobasepairs) anneals to a denatured target DNA (genomic DNA or PCR amplification product) with sequence homology. Target DNA can be directly blotted onto a membrane, or if size information of the hybridization target is warranted, the target DNA is first run through agarose gel and then transferred to a membrane. Detection of hybrids is based on a radioactive signal, fluorescence, or color reaction, depending on the type of the label. By determining the intensity of the hybridization signal, the number of target organisms can be estimated (Felske, 1999). With dot-blot hybridization, nucleic acids can be fairly rapidly analyzed for the presence of specific sequences. This technique is commonly used to confirm the identity of PCR products (Lipski et al., 2001).

2.3.6 Fluorescent in situ hybridization

The detection of whole-bacterial cells via labeling of specific nucleic acids with fluorescence-labeled oligonucleotide probes is called fluorescent in situ hybridization (FISH). The advent of nucleic acids technologies has provided a means by which microbial communities can be studied and classified without the limitations of cultivation. Fluorescent in situ hybridization (FISH) is now being used to investigate community structure and diversity of soils, aquifers, and other natural communities (Rogers et al., 2007). The whole-cell or in situ hybridization technique is now a much-used molecular tool in environmental microbiology, since organisms or groups of organisms can be identified with minimal disturbance of their environment and spatial distribution. It remains a key technique in microbial ecology (Lenaerts et al., 2007). It is a very useful technique for detection of targeted chromosomal abnormalities (Xi et al., 2003). FISH in combination with epifluorescence microscopy is a widely applied method to analyze microbial communities. The sensitivity and objectivity can be greatly enhanced by digital image analysis.
2.3.7 Genetic fingerprinting techniques

Genetic fingerprinting techniques can be used to characterize bacterial communities or single bacterial isolates. The genetic fingerprinting of microbial communities provides a pattern or profile of the community diversity, based upon the physical separation of unique nucleic acid sequences (Meyer et al., 2007). Community analysis techniques are relatively easy and rapid to perform and they allow simultaneous analysis of multiple samples, enabling the comparison of the genetic diversity of microbial communities from different habitats, or the study of the behavior of individual communities over time. Community analysis can be performed with techniques such as denaturing-gradient gel electrophoresis (DGGE), temperature-gradient gel electrophoresis (TGGE), and single-stranded conformational polymorphism (SSCP). There is also an approach based on heteroduplex mobility analysis of 16S rDNA fragments for targeted detection of sub-populations of bacteria within diverse microbial communities (Turner et al., 2002). Fingerprinting of bacterial isolates can be performed by a variety of techniques, including e.g. ribotyping, amplified ribosomal DNA restriction analysis (ARDRA), pulsed-field gel electrophoresis (PFGE), RAPD, repetitive element sequence-based PCR (rep-PCR), and amplified fragment length polymorphism (AFLP). All these techniques aim at differentiating bacterial isolates at the subspecies level, preferably even at the strain-level.

2.3.8 Denaturing/thermal gradient gel electrophoresis (DGGE/TGGE)

In DGGE and TGGE, PCR-amplified DNA fragments of the same length but with different DNA sequences can be differentiated. This method has recently been applied to the analysis of 16S rRNA genes from environmental samples and allows the separation of a heterogeneous mixture of PCR amplified genes on a polyacrylamide gel (Madigan et al., 2003). Separation in DNA fragments is based on the electrophoretic mobility of a partially melted double-stranded DNA molecule in polyacrylamide gels containing either a linear gradient of DNA denaturants (a mixture of urea and formamide in DGGE) or a linear temperature gradient (TGGE). Partially melted DNA fragments are held together with a G+C-rich oligonucleotide, a GC-clamp. Therefore, each denaturing fragment generates only a single band in the gel. DGGE/TGGE performed after PCR gives an
insight into the predominant microbial populations; and DGGE/TGGE performed after reverse transcriptase (RT)-PCR helps identify the predominant active microbial populations (Duineveld et al., 2001). DGGE/TGGE can also be used in combination with quantitative RT-PCR to quantify rRNA sequences in complex bacterial communities. DGGE/TGGE analysis combines a direct visualization of bacterial diversity and the opportunity to subsequently identify community members by DNA fragment sequence analysis or hybridization with specific probes. DGGE/TGGE has some specific limitations. DGGE/TGGE can be used to separate only relatively small fragments and it displays only the rDNA amplicons obtained from the predominant (over 1% of the population) species present in the community. The presence of heterogeneous 16S rRNA genes (16S rRNA genes that exhibit small sequence variations in the genome of a given strain) can result in several bands in a DGGE/TGGE profile. Furthermore, a single band may represent more than one strain. The construction of 16S rDNA clone libraries and the screening for different clones by DGGE may overcome these deficiencies (Schabereiter-Gurtner et al., 2001). DGGE fingerprinting does not provide complete coverage of bacterial diversity since only a limited number of bands can be resolved in any fingerprint (Edenborn and Sexstone, 2007).

2.3.9 Single-stranded conformation polymorphism (SSCP)

Single-strand conformational polymorphism (SSCP) analysis, is widely accepted approaches to describe bacterial diversity based on nuclear ribosomal DNA genes and internal transcribed spacer (ITS) regions (Van Oppen, 2007). SSCP analysis detects sequence variations between different DNA fragments, which are usually PCR-amplified from variable regions of the 16S rRNA gene. The technique is based on the fact that a single base modification can change the conformation of a single-strand DNA molecule, altering the migration speed of the molecules in a non-denaturing gel (Ouverney and Fuhrman, 1999). DNA fragments of the same size but with different base composition can thus be separated. The limitations of the SSCP method are discriminatory power and reproducibility of SSCP analysis is usually most effective for fragments up to 400 bp in size, depending on the length of the fragment studied, the position of the sequence variations in the gene studied, and the test conditions. In addition, PCR-SSCP detects
bacterial populations that make up 1% or more of a bacterial community. A major limitation of SSCP for community analysis is the high rate of DNA strand-annealing after the initial denaturation during electrophoresis. Besides community studies, PCR-SSCP analysis can be adapted for the rapid identification of Gram-negative and Gram-positive bacteria at the genus and species levels (Wagner et al., 2000).

### 2.3.10 Ribotyping

Ribotyping, also referred to as ‘molecular fingerprinting’, is a way of identifying microorganisms from the analysis of DNA fragments generated from genes encoding their 16S rRNA (Hartel et al., 2002; Samadpour, 2002). The ribotyping procedure provides a DNA fingerprint of bacterial genes coding for rRNA, which are highly conserved in microorganisms. The genetic fingerprints of the bacterial isolates from the different samples can then be compared. The ribotyping procedure follows: bulk DNA or DNA encoding 16S rRNA and related genes within the rRNA operon is PCR amplified and separated by electrophoresis. The pattern is generated from the DNA fragments on the gel which is used to make comparisons of the patterns with references in the database. The unique pattern of DNA bands are used to determine the closest database match of the bacteria. Ribotyping is one of the most rapid and specific method of bacterial identification, used world wide (Scott et al., 2003).

### 2.3.11 Amplified rDNA restriction analysis

In addition to classic ribotyping, rDNA-based fingerprints can be obtained by a technique called ARDRA. The ARDRA technique using only one PCR reaction and one restriction enzyme can discriminate among bacteria at genus levels (Collado and Hernandez, 2007). In ARDRA, bacterial rRNA gene(s) are first amplified by PCR, using conserved sequences of rDNA as primers. The PCR amplification product is then digested with restriction endonuclease and restriction fragments are resolved electrophoretically to obtain a fingerprint (Vaneechoutte et al., 1992). Although ARDRA fingerprinting is faster to perform than classic ribotyping, its discriminatory power is often inferior to that of ribotyping. This is due to the fact that smaller areas of the rRNA
operon (and none of the sequences surrounding the rRNA genes) are targeted in ARDRA than in ribotyping.

### 2.3.12 Pulse field gel electrophoresis (PFGE)

Pulsed-field gel electrophoresis (PFGE) has been the standard technique for strain typing most bacterial species (Ross and Heuzenroeder, 2005). The chromosome is the most fundamental component of identity of the cell and therefore represents a preferred measure for assessing strain interrelatedness. Restriction enzymes that cleave chromosomal DNA less frequently have been utilized for analysis. The resulting DNA fragments are too large to be separated by conventional agarose gel electrophoresis. A number of alternative methods, generally classified as PFGE, are capable of separating these large DNA fragments (Chang and Chui, 1998). In conventional agarose gel electrophoresis, DNA molecules that are more than 40 to 50 kb in size fail to migrate efficiently. By periodically changing the direction of the electrical field in which the DNA is separated, PFGE allows the separation of DNA molecules of over 1,000 kbp in length (often referred to as megabase-sized DNA). PFGE methods differ in the way the pulsed electric field is delivered to the agarose gel. To interpret DNA fragment patterns generated by PFGE and transform them into useful information microbiologist must understand how to compare PFGE patterns and how random genetic events can alter these patterns.

### 2.3.13 Random amplified polymorphic DNA (RAPD)

RAPD is frequently used to characterize probiotic bacteria. In RAPD fingerprinting, one or two primers (usually 10–12 bp long) are arbitrarily selected and allowed to anneal to the bacterial genomic DNA template at a low stringency. Due to its technical simplicity and speed, RAPD methodology has been used for diversity analyses (Iqbal et al., 2007). In RAPD, several amplification products of varying sizes are obtained. These products are resolved electrophoretically to yield a RAPD-fingerprint. RAPD typing is fast to perform, especially in cases where fingerprinting can be performed directly on single colonies growing on an agar plate. Due to the low stringency of the PCR amplification, RAPD-fingerprints can show some variation (especially in
band strengths) and therefore the fingerprint comparisons have to be done visually by an experienced person. However, when strictly identical conditions (same thermocycler, reagents, etc.) are used, the method usually works well. RAPD banding-pattern reproducibility can be improved by using a procedure where the same strains are exposed to three different annealing temperatures (with increasing stringency) and by identifying the stable amplicons (Cusick and Sullivan, 2000). This triplicate procedure naturally makes the RAPD fingerprinting technique more laborious.

### 2.3.14 Repetitive element sequence-based PCR (rep-PCR)

Repetitive element sequence-based PCR (rep-PCR) is a typing method that differentiates microbes by using primers complementary to interspersed repetitive consensus sequences that enable amplification of diverse-sized DNA fragments consisting of sequences between the repetitive elements (Spigaglia and Mastrantonio, 2003). Repetitive chromosomal elements, which are found randomly distributed in bacterial genomes, are the targets of rep-PCR amplification. In rep-PCR, primers anneal to repetitive parts of the chromosome and amplification occurs when the distance between primer binding sites is short enough to enable this. The repetitive DNAs can be classified either as short sequence repeats (SSRs) or variable number of tandem repeats (VNTRs). Variations of rep-PCR include enterobacterial repetitive intergenic consensus PCR (ERIC-PCR), BOX-PCR, repetitive extragenic palindromic unit sequence PCR, and VNTR-PCR (Van Belkum, 1998).

### 2.3.15 Amplified fragment length polymorphism (AFLP)

The amplified fragment length polymorphism (AFLP) technique is being increasingly used in phylogenetic studies (Althoff et al., 2007). AFLP is a typing method that utilizes a combination of restriction enzyme digestion and PCR (Whatmore et al., 2005). In the AFLP procedure, the DNA is digested with two different restriction endonucleases, usually chosen so that one cuts more frequently than the other. This restriction strategy generates a large number of fragments. In order to make the interpretation of the results more feasible, only a specific subset is used for isolate comparison. The subset is generated by linking adapter sequences to the ends of the restriction fragments extending the length of the known end sequences. PCR primers are
designed to hybridize to the adapter sequence, the remaining restriction site sequence, and an additional one or two nucleotides of the unknown template sequence. The addition of each nucleotide, chosen at random, to the end of the primer reduces the number of fragments that will be amplified by a factor of four. Following PCR, the reaction products are separated by gel electrophoresis and their banding patterns are resolved. The method utilizes the benefits of RFLP analysis with the increased sensitivity of PCR to generate profiles that are reproducible and relatively easy to interpret and compare to those for other isolates from a nosocomial outbreak by using a software program such as BioNumerics. Although the AFLP technique offers many advantages for generating phylogenetic markers, the phylogenetic utility of the markers is limited by the assumption of size homology—identifying the homology of the fragments is not simple (Bussell et al., 2005).

A comparison of the above described culture dependent methods is shown below (Fig. 2.1).

Fig. 2.1: Comparison of the procedural steps involved in various molecular typing methods (Olive and Bean, 1999)
Table 2.1: Overview of genetic fingerprinting techniques

<table>
<thead>
<tr>
<th>Technique</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGGE/TGGE and SSCE</td>
<td>Community structure and dynamics can be studied. Identification of community members possible</td>
<td>Only those populations making up over 1% of the total community can be detected</td>
</tr>
<tr>
<td>Strain level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ribotyping</td>
<td>Can be automated, Good discriminatory power, Can be used for bacterial identification</td>
<td>Expensive, laborious, and manually slow to perform</td>
</tr>
<tr>
<td>ARDRA</td>
<td>Fairly simple and fast</td>
<td>Limited discriminatory power</td>
</tr>
<tr>
<td>PFGE</td>
<td>Very high discriminatory power</td>
<td>Expensive, Slow to perform</td>
</tr>
<tr>
<td>RAPD</td>
<td>Fast, simple, and cost-effective</td>
<td>Reproducibility problems possible</td>
</tr>
<tr>
<td>rep-PCR</td>
<td>Fast, simple, and cost-effective</td>
<td>Reproducibility problems possible</td>
</tr>
<tr>
<td>AFLP</td>
<td>Good discriminative power</td>
<td>Expensive, Laborious</td>
</tr>
</tbody>
</table>

2.3.16 Limitations of molecular microbial ecology

We have gained new and valuable knowledge using the techniques described above, but there are important limitations that must be minimized, or at least, recognized. The limitations related to the extraction of nucleic acids from natural samples, artifacts associated with enzymatic amplification, cloning of PCR products, primers pairing etc have been described.

Nucleic acid extraction

A major limitation of all the methods described is the quantitative recovery of DNA from environmental samples. There is always the argument that if we do not know the total amount of nucleic acids present in a sample, then it is difficult to assess the efficiency of recovery by any extraction technique. Gram-positive cells are more resistant to cell lysis than Gram-negative cells. Efficiency of cell lysis in an environmental sample can be obtained by microscopic enumeration of the cells in a sample before and after lysis treatments. There are many methods for extracting DNA from natural samples, but there have been few systematic studies that have addressed this issue (Padmanabhan et al., 1998; Lueders et al., 2003).
It is possible that the same lysis technique may give different results with different types of sample such as water, sediment, or soil, and the degree of cell lysis should be determined independently. It has been demonstrated that a combination of physical and chemical treatments, such as freezing and thawing, lysis with detergents, and bead beating lysed approximately ninety-six percent of cells in soil with high efficiency. However, smaller cells (0.3-1.2 \( \mu \text{m} \)) are more resistant to lysis. This has implications for recovery of sequences from environmental samples where many cells are likely to be small. It has been found, that even without harsh physical treatments such as bead beating, up to 99.8% lysis can be obtained, although this did require long incubations with lysozyme and up to six freeze thaw cycles (Von et al., 2001).

**PCR and cloning**

Selectivity in PCR amplification of rRNA genes can affect the results of molecular biological measures of diversity. Small differences in the sequence of universally conserved regions may result in selective amplification of some sequences, particularly when primer annealing is at high stringency. The frequency of different sequence types in PCR-derived rRNA gene clone libraries represents the relative abundance of different components of a microbial community. The copy number of the genes in each genome and the size of the genome of bacteria are known, and this can be accounted for estimation of species abundance. In natural samples, more abundant sequences are preferentially amplified whereas, low-abundance sequences are discriminated and there is no such information about the constituent microbial types. It has been observed that high \% G+C templates are discriminated due to lower efficiency of strand separation during the denaturation step of the PCR (Pandey et al., 2007). PCR amplification using mixtures of genomic DNA from organisms with different genome sizes and numbers of rRNA operons has shown that, the ratio of rRNA genes in the PCR product mix reflect the ratio in the starting mixture of DNA (Buzdin et al., 2007).

**Primer pairing**

It has been demonstrated that some primer pairs gave a strong correlation between the ratio of genes in the starting mix and the ratio in the final PCR product, but in
environmental samples, this may not be a problem. However, the presence of closely related sequences from different taxa was not considered, and it would be required to resolve this problem (Suzuki and Giovannoni, 1996). This also has been observation that cloned PCR products generated using different primers resulted in significantly different composition of clone libraries. Furthermore, it has been found that the same batch of PCR product cloned using either blunt-end or sticky-end cloning procedures gave different results (Pandey et al., 2007).

**Enzymes**

The fidelity of PCR amplification varies, depending on the particular thermostable DNA polymerase used (manufacturers have quoted misincorporation rates in the range of 0.000002%–1.3% for different thermostable DNA polymerases). Careful analysis of secondary interactions should identify discrepancies due to misincorporation of nucleotides during PCR. Furthermore, inter operon differences of up to 5% in rRNA gene sequence have been reported. This degree of sequence divergence can be associated with rRNA sequences from different individual species, as well as intra-operon variability (Richard et al., 2002; Satyanarayana et al., 2006). This is of much concern when making conclusions about biodiversity data obtained from rRNA gene clone libraries. This degree of sequence divergence is associated with different, described species and even genera in some lineages.

**Chimeric PCR products**

The formation of chimeric PCR products has been observed in which fragments from two different sequences become fused during the amplification process. One study demonstrated that up to 30% of products generated during co-amplification of similar templates were chimeric. Nonetheless, the results demonstrated the considerable potential for chimera formation during PCR amplification. A number of computer programs have been developed to help identify chimeric sequences, but these have difficulty in identifying chimeras when the two sequences from which the chimera is formed show greater than 85% homology. The programs are best used as a guide to the presence of chimeric sequences. The authenticity of a sequence should be confirmed by independent
sequence analyses, using the putative chimeric fragments. Discrepancies in the secondary structures also aid in the identification of chimeric molecules (Hugenholtz and Huber, 2003; Gonzalez et al., 2005).

2.4 Exploring prokaryotic diversity

Microorganisms isolated in pure culture from an environment represent the numerically dominant or functionally significant species in that environment. In fact, microorganisms isolated using standard cultivation methods are rarely numerically dominant in the communities from which they were obtained. They are isolated by virtue of their ability to grow rapidly into colonies on high-nutrient growth media, under laboratory conditions, at moderate temperatures and thus represent the dominantly cultivated population. Easily isolated organisms are called 'weeds' of the microbial world and are estimated to constitute less than 1% of all microbial species. Recent researches have shown that there is a largely unexplored microbial diversity in extreme environments, such as hot springs, hydrothermal vent sites, ocean and sea ice, hypersaline environments and environments exhibiting extremes of pH. Some of the environments are described here (Keller and Zengler, 2004).

2.4.1 Extreme environments

Ocean environment

Certain areas of the ocean seem to be dominated by particular groups of microorganisms. Members of the archaeal phylum Crenarchaeota are shown to comprise 39% of detected microbial phyla at certain depths. In the ocean, often referred as a floating desert, (because of its very poor levels of nutrients), how could such a rich diversity of species survives. In certain regions, SAR11 comprise up to 50% of the total surface bacterial community and approximately 25% of the sub-euphotic zones. On average, members of the SAR11 clade account for about one third of all cells in surface waters. So far, only one representative of this clade has been isolated (Massana et al., 2000). Both the great abundance and global distribution of SAR11 suggests that its members are instrumental in metabolizing oceanic dissolved organic matter, but the specific roles of these bacteria in biogeochemical cycles is still unknown (Malmstrom et
al., 2004; 2005). We now know that the highest percentage of bacterial 16S ribosomal genes present in all oceanic and coastal waters are from members of the SAR11 clade, making this group “one of the most successful clades of organisms on the planet” (Giovannoni et al., 2005).

**Soil environment**

Even though the open ocean is stratified and heterogeneous, soil is much more diverse environment, and this is reflected in the diversity of the physiological and metabolic capabilities of the soil biota (MacNeil et al., 2001). According to an estimate, pasture soil sample contains about 3,500-8,800 genome equivalents. This could result in approximately 10,000 different species of equivalent abundance. This immense diversity presents a challenge to describe microbial communities (Daniel, 2004). Even small changes in the soil environment seem to have a large influence on the overall species diversity and community structure. Some of the environmental factors that are known to influence community composition are heavy metals, organic contaminants pesticides and physical parameters.

**Subsurface water communities**

In the microbial diversity of subsurface and deep water, samples include the resident active microbial community and cells that are deposited or transported into the subsurface environment. Studies of metabolic activity in oceanic and terrestrial subsurface, as well as substrate availability, provides understanding of microbial life in subsurface environments (Kopke et al., 2005; Polz et al., 2006). In some cases, subsurface seems to be a preferred environment for microorganisms that thrive on hydrogen. However, subsurface environments and marine sediments are as diverse as deep ocean environments, and thus resulting in different microbial communities (Hondt et al., 2004; Brito et al., 2006).

**Hot springs**

Many hyperthermophilic archaea and bacteria have been isolated from hot spring environments. On the basis of environmental studies and isolation of similar archaeal
phytotypes from other hydrothermal environment, these organisms were believed to be
the dominating microorganisms in hot springs. A study of microbial diversity in hot
spring samples showed that although all the hot springs in close geographical proximity
had similar temperatures (between 85°C and 95°C) and a comparable pH value (7.8-8.9),
but they differed markedly with regard to their overall microbial diversity (Blank et al.,
2002). This finding indicated that, in common with the complexity of microbial diversity
in other environments, geochemical variations affect biodiversity. Studies that
incorporate parameters like environment specific sampling will allow us to understand
total biodiversity.

**Hydrothermal vents**

Hydrothermal vents occur over a wide depth range, from intertidal to the abyss
(Tarasov et al., 2005). Since the discovery the deep-sea hydrothermal vent in the 1970s,
these environments have been regarded as one of the main habitats for thermophiles.
Indeed, many thermophilic archaeal and bacterial species have been isolated from deep-
sea vent sites. Molecular analyses of the hydrothermal vent communities has revealed
phytotypes that have not yet been found in cultivated or in 16S rRNA clone libraries. To
understand the microbial diversity in these dynamic deep-sea environments, sampling
procedures and measuring parameters are crucial (Blank et al., 2002).

**Polar ice caps**

The discovery of icy bodies in our Solar System has opened up the possibility that
life may exist on and below their surfaces. Snow and ice are good sites for the
preservation of biomarkers as they trap and preserve organic matter that is deposited onto
their surfaces (Phillips and Parnell, 2006). The ice of the polar oceans can cover up to
13% of the earth’s surface area. This environment is a fertile habitat for bacteria,
microscopic plants and animals. 16S rRNA gene sequencing indicated that the
community composition in samples of sea ice was linked to bacterial phytotypes,
recovered from heterotrophic bacterial taxa (Thomas and Dieckmann, 2002). They
included several new genera and species.


**Hypersaline habitats**

Hypersaline environments, which contain almost saturating concentrations of sodium chloride, are often dominated by archaeal and bacterial halophilic microorganisms (Benlloch et al., 2002). However, the overall diversity in extreme hypersaline environments is generally low. Besides the dominance of archaea, a new bacterium, *Salinibacter ruber*, that is an important component of the halophilic microbial communities. *S. ruber* can comprise up to 25% of the total prokaryotic community in hypersaline environments (Alio, 2006a).

**Extreme pH environments**

The overall microbial diversity in extreme pH environments is usually low compared with microbial diversity in non-extreme pH environments. Molecular studies have revealed that *Ferroplasma* species constitute up to 85% of the microbial community in extremely low pH environment. Alkaline environments, such as soda lakes, have elevated temperatures and high salinity and therefore populated by microorganisms similar to found in hypersaline environments. Very few natural alkaline environments, which lack increased levels of salinity and temperature, have been found. Half of the bacterial phylotypes obtained from these environments form a specialized and unique psychrophilic community with sequences less than 90% similarity to sequences in the available databases (Edwards et al., 2000). Many acidophiles are sources of enzymes like amylases, glucoamylases, proteases, cellulases, oxidases (Van den Burg, 2003). An analysis of the microbial diversity in combination with isolation parameters, has allowed linking of community composition with environmental factors and microbial diversity lying within these environments (Keller and Zengler, 2004).

**2.4.2 Improving strategies for cultivation**

Cultivation techniques are improving and have allowed to grow in vitro an increasing number of still uncultured microorganisms, our knowledge of their growth conditions in nature (i.e., chemistry of the original environment, life in complex communities, obligate interactions with other organisms, etc.) remains insufficient to cultivate most of them (Eyers et al., 2004). In fact, most of the species in many
environments have never been described, and this situation will not change until new culture technologies are developed. It is expected that one of the largest efforts in this field of research in the next decade will be the exploration of a means to grow less-culturable marine microbes. It is thought that the reason for the enormous discrepancy between the total viable cell count and the cell count of culturable cells might be because of the following: cell damage by oxidative stress; formation of viable but non-culturable cells; inhibition by high substrate concentrations; induction of lysogenic phages upon starvation; and lack of cell–cell communication in laboratory media. Among the approaches used to enhance the resuscitation of less culturable strains, are the addition of cell signaling molecules and the use of oligotrophic isolation media (Bruns et al., 2002; Connon and Giovannoni, 2002).

Only 27 out of 53 bacterial phyla contain previously cultivated microorganisms, with many phyla represented by only a few isolates and some phyla containing only one described species (Rappe and Giovannoni, 2003). The Acidobacteria and Verrucomicrobia divisions are among those divisions of the domain Bacteria represented by a large diversity of 16S rRNA genes, which occur in particular abundance in soils, but contain few cultured members (Buckley and Schmidt, 2003). Hence, our appreciation of the physiological diversity of Acidobacteria and Verrucomicrobia is limited, as is our knowledge of their role in global biogeochemical cycles. Clearly, a better understanding of these divisions would be attained by having a greater diversity of their members available in pure culture for detailed study. The intrinsic selectivity of any given medium and incubation condition imposes limits on the nature, number, and diversity of microbes recovered from natural samples. It follows, then, that the application of isolation procedures that better mimic conditions existing in the habitat from which the samples were obtained could increase the likelihood of retrieving previously uncultured organisms. Efforts to accomplish this have met with some success by using the following: (i) relatively low concentrations of nutrients (typically three orders of magnitude less than common laboratory media). Many novel strains among the Proteobacteria, Planctomycetes, Bacteroidetes, Acidobacteria and Verrucomicrobia were obtained; these included some previously believed to be uncultivable (Connon and Giovannoni, 2002), (ii) nontraditional sources of nutrients, signaling molecules, or inhibitors (of undesired
organisms) (Bruns et al., 2002); and (iii) relatively lengthy periods of incubation (Connon and Giovannoni, 2002; Joseph et al., 2003; Rappe and Giovannoni, 2003), sometimes directly in the natural environment from which the inoculum was obtained (Kaeberlein et al., 2002). Whatever cultivation approach is tried, however, one is ultimately confronted with the need to evaluate its success. This is a potentially arduous task if, many different media and incubation conditions are being tested and little or nothing is known about the microbes sought other than their 16S rRNA gene sequences. Accordingly, some high-throughput screening method is desirable. To deal with this, a simple, high-throughput, PCR-based procedure, plate wash PCR (PWPCR) was developed that facilitated the surveillance of isolation plates for the presence of target organisms and the ultimate recognition of colonies made up of target organisms (Fig 2.2) (Stevenson et al., 2004).

Fig. 2.2: PWPCR method to detect growth and monitor isolation of targeted bacteria. Of the three medium and incubation conditions shown in this diagram (conditions A, B, and C), growth of targeted bacteria (+) is represented only in condition C (Stevenson et al., 2004).
A high throughput cultivation method based on the combination of single cell encapsulation and a gel microdroplet was also developed. Gel microdroplets are small porous gel matrixes that encapsulate individual cells. This method allows the exchange of metabolites and/or signaling molecules that are produced by other microorganisms. It enables cells to grow with nutrients that are present in their natural environment. After long periods of incubation with a constant flow of a very low concentration of nutrients, flow cytometry was used to separate gel beads containing cells that had grown into small microcolonies from those that had not. Many novel bacteria have been isolated in this way, including members of previously uncultured groups that are abundant in seawater (Keller and Zengler et al., 2004).

2.5 Some key discoveries revealed by the application of the rRNA approach to studies of microbial diversity and ecology

Despite its limitations, this technology is permitting major advances in our understanding of microbial ecology and evolution. Its potential lies not only in the identification of specific organisms in the environment, but also in its ability to complement other methods (including classical microbiology and process-related studies), to assign them to functional roles, and to assess their significance in environmental processes. This section considers some examples of systems, previously refractory to classical techniques alone, where molecular analysis has played an expanding role.

Microbial ecology of morphologically distinct bacteria important in wastewater treatment

Activated sewage sludge contains a considerable diversity of microorganisms, filamentous bacteria in sewage sludge have long been defined morphologically with little or no information on the physiology or evolutionary relationships of the bacteria observed. A number of these filamentous forms have been associated with operational problems in sewage treatment plants, such as sludge bulking e.g., Thiothrix spp. (Payne et al., 2007) and Beggiatoa spp. (Galvan and de Castro., 2007) and foaming e.g., Microthrix
parvicella (Tsai et al., 2003) that cause problems with solids separation. However, a number of these filamentous species are difficult to distinguish, even morphologically, and some are known to exhibit nonfilamentous growth. A suite of probes that could distinguish Haliscomenobacter hydrossis (Martin and Schauer, 2007), Thiothrix nivea and Leucothrix mucor (Payne et al., 2007). Eikelbloom type 021N, some strains of “Leptothrix discophora (Nakatsu et al., 2006) and Sphaerotilus natans (Galvan and de Castro, 2007) were used to analyze the bacterial communities of many activated sludge plants.

T. nivea was, filamentous type and is responsible for sludge bulking “Microthrix parvicella” is a filamentous Gram-positive bacterium associated with activated sludge bulking and foaming. “Microthrix” is notoriously difficult to maintain in pure culture, and little information has been obtained regarding the physiology and taxonomy of the bacterium (Levantesi et al., 2006). However, pure cultures of “M. parvicella” have been obtained by painstaking micromanipulation of individual filaments onto different growth media. After 5 weeks of incubation, colonies were just visible on agar plates. Sufficient DNA allow 16S rRNA genes to be amplified and sequenced. Phylogenetic analyses placed “M. parvicella” with Actinomycetes in the high % G+C Gram-positive bacteria.

Ammonia-oxidizing bacteria

Autotrophic ammonia-oxidizing bacteria are an ecologically important and physiologically specialized group. They are responsible for the oxidation of ammonia to nitrite, the reaction that drives the process of nitrification in a wide range of environments. The study of autotrophic ammonia oxidizing bacteria in natural environments using culture based methods is problematic. Molecular biological methods provide a useful alternative approach. Autotrophic ammonia- oxidizers are notoriously difficult to isolate in pure culture from environmental samples (Hidetoshi et al., 2006). Studies on the physiology of ammonia oxidizing bacteria have been almost exclusively centered on Nitrosomonas europaea, because of the comparative ease with which this species can be grown in culture. This has resulted in ammonia oxidation in the environment often being equated with the activity of nitrosomonads (Tarre et al., 2007).
Morphologically conspicuous but as yet uncultivated bacteria

Magnetotactic bacteria

Magnetotactic bacteria are a heterogeneous group of aquatic prokaryotes with a unique intracellular organelle, the magnetosome, which orients the cell along magnetic field lines (Richter et al., 2007). Magnetotactic bacteria were first described over two decades ago and were perhaps considered microbial oddities. Subsequent to their original description, we now know that they are common in many freshwater and marine sediments. In some environments, a single species of magnetotactic bacterium has been shown to be the dominant component of the bacterial population (e.g., the microoxic zone of a German freshwater lake). The difficulty in cultivating magnetotactic bacteria has hampered progress in the understanding of these interesting organisms. However, magnetic enrichment procedures have been used to obtain relatively purified preparations of magnetotactic bacteria from freshwater sediments. These were used to obtain 16S rRNA gene sequences from which fluorescently-labeled oligonucleotide probes were designed and used to determine what cells were the source of the sequences obtained. This revealed the presence of three morphologically similar but phylogenetically distinct magnetic cocci. All of the magnetotactic cocci belonged to the $\alpha$-Proteobacteria. The first report of a cultivated magnetic coccus was published in 1993 (DeLong et al., 1993).

Achromatium oxaliferum

Achromatium oxaliferum is a remarkable bacterium. It can be greater than 100 μm in length and up to 30 μm in diameter. It deposits intracellular sulfur and is unique among the Bacteria in precipitating intracellular calcite inclusions (Gray et al., 2007). The paucity of information about A. oxaliferum is due, in large part, to the continuing inability to cultivate the bacterium. Fortunately, cells of A. oxaliferum can be readily purified from crudely screened sediments by virtue of their high specific gravity conferred by the intracellular calcite inclusions. This has allowed sufficient purified DNA to be extracted to allow PCR amplification and sequencing of 16S rRNA genes (Head et al., 1996). Direct sequencing of PCR products amplified from A. oxaliferum DNA gave poor sequence data, but cloning of the PCR product and sequencing a number of the clones
revealed substantial heterogeneity in the sequences obtained from purified cells. Three major lineages were apparent from the eight clones partially sequenced. The cloned sequences, however, formed a strongly supported monophyletic group within the $\gamma$-Proteobacteria. The subdivision of the $\gamma$-Proteobacteria within which the *A. oxaliferum* sequences fall is dominated by sulfide-oxidizing autotrophs.

**Epulopiscium fishelsoni**

The gut of a number of species of surgeonfish (family Acanthuridae) are known to harbor large (up to 80 × 600 mm) symbiotic microorganisms. These were originally described as eukaryotic protists because of their large size (Fishelson et al., 1985). The ultrastructure of the bacterium, as determined by electron microscopy, was, however, more characteristic of prokaryotic organisms. Like many of the magnetotactic bacteria, *E. fishelsoni* remains uncultured, so taxonomic inferences based on the biochemistry or physiology of the organism have been difficult to establish. Angert et al. (1993) used the PCR to obtain 16S rRNA sequences from *E. fishelsoni* cells purified by micromanipulation. Comparative sequence analyses demonstrated that three sequence types obtained from *E. fishelsoni* purified from Australian surgeonfish formed a monophyletic group related to *Clostridium* spp. in the low % G+C Gram-positive bacteria.

### 2.6 Framework for mapping biodiversity

Comparative analysis of small-subunit ribosomal RNAs provided an objective framework for determining evolutionary relationships between organisms and thereby 'quantifying' diversity as sequence divergence on a phylogenetic tree. Cellular life divided into three primary lineages: one eukaryotic and two prokaryotic (Bacteria and Archaea), and 11 major lineages are also defined (phyla or divisions) within the bacterial domain on the basis of 16S rRNA sequences obtained from cultivated organisms (Woese, 1987). This revealed distant relationships which were not suspected from phenotypic characterization. The most intensively studied prokaryotic genera are members of only three bacterial phyla: Proteobacteria, Firmicutes and Actinobacteria.
Chapter 2

Review of Literature

2.7 World of uncultivable microbes

Microbiology has experienced a transformation during the last 25 years that has altered microbiologists view of microorganisms and how to study them. The realization that most microorganisms cannot be grown readily in pure culture forced microbiologists to question their belief that the microbial world had been conquered (Handelsman, 2004). This has revolutionized the microbiological thought and scientists started believing that the uncultured microbial world far outsized the cultured world and that this unseen world could be studied. This change in thinking was prompted by another, equally important realization: microorganisms underpin most of the geochemical cycles and many human health conditions that were previously thought to be driven by inorganic processes and stress, respectively. The glimmers of insight into the influence that microorganisms exert on the world propelled microbiologists to pursue the uncultured world.

2.7.1 A pure culture is not enough

Because culturing provided the platform for building the depth and detail of modern microbiological knowledge, for a long time microbiologists ignored the challenge to identify and characterize uncultured organisms. They focused instead on the rich source of discovery found in the readily cultured model organisms, and this contributed to the explosion of knowledge in microbial physiology and genetics in the 1960s to mid-1980s. One of the indicators that cultured microorganisms did not represent much of the microbial world was the oft-observed “great plate count anomaly” (Staley and Konopka, 1985)—the discrepancy between the sizes of populations estimated by dilution plating and by microscopy.

Further evidence that drew attention to the uncultured world accumulated during the 1970s and 1980s. A study of oligotrophs indicated that incubation times longer than 25 days enhanced the recovery of certain organisms in culture (Whang and Hattori, 1988). The confluence of these and many other scientific and technical advances steadily drew attention to the unculturable microbial world, but two discoveries figured significantly in the sharpened focus. The first was work on the diversity of soil bacteria, which demonstrated with DNA-DNA reassociation techniques that the complexity of the bacterial DNA in the soil was at least 100-fold greater than could be accounted for by
culturing. The second discovery was the demonstration that *Helicobacter pylori* causes gastric ulcers and cancer. These discoveries provided compelling evidence that drew microbiologists to wrestle with the daunting challenge of devising strategies to access these organisms.

### 2.7.2 The paradigm shift

In 1985, an experimental advance radically changed the way we visualize the microbial world. Building upon the pioneering work of Carl Woese, which showed that rRNA genes provide evolutionary chronometers (Woese, 1987), Pace and colleagues created a new branch of microbial ecology. They used direct analysis of 5S and 16S rRNA gene sequences in the environment to describe the diversity of microorganisms in an environmental sample without culturing. The next technical breakthrough arrived with the development of PCR technology and the design of primers that can be used to amplify almost the entire gene (Giovannoni et al., 1990). The application of PCR technology provided a view of microbial diversity that was not distorted by the culturing bias and revealed that the uncultured majority is highly diverse and contains members that diverge deeply from the readily culturable minority. Till date, 53 phyla have been delineated, and most are dominated by uncultured organisms (Fig. 2.3) (Rappe and Giovannoni, 2003).

Large-scale genomic surveys of microbial communities are currently expanding massively in number, scope and pace. Recent genomic forays into complex microbial communities include acid-mine drainage sites (Tyson et al., 2004), symbiotic associations (Woyke et al., 2006), pollutant removing bioreactors (Martin et al., 2006) and the human microbiome (Turnbaugh et al., 2006). As microbial community genomic surveys accumulate, deciphering the genetic and functional “differences that make a difference” within and between different microbial habitats is becoming ever more feasible (DeLong., 2007). Craig Venter and colleagues reported a (mostly) ocean surface water microbial sequencing survey that has nearly doubled the number of known protein sequences (Rusch et al., 2007). The 41 randomly collected microbial samples in the ‘Global Ocean Sampling’ (GOS) cumulatively encompass ~6.6 billion base pairs of DNA, translating into ~6 million predicted protein sequences (Yooseph et al., 2007).
Fig. 2.3: Phylogenetic tree of Bacteria showing established phyla and candidate phyla. The vertex angle of each wedge indicates the relative abundance of sequences in each phylum; the length of each side of the wedge indicates the range of branching depth found in that phylum; the redness of each wedge corresponds to the proportion of sequences in that phylum obtained from cultured representatives. Candidate phyla do not contain any cultured members (Handelsman, 2004).
Only 27 of the approximately 53 identifiable major lineages, or phyla, within the
domain Bacteria have cultivated representatives. Evidence from field studies indicates
that many of the cultivated phyla are found are found in diverse habitats, and some are
extraordinarily abundant (Rappe and Giovannoni, 2003). The lineages that do not have
any culturable representative are known as candidate phyla or divisions, with the term
candidate implying that no cultures yet exist to represent the group.

2.7.3 Metagenomics

Metagenomics, the genomic analysis of a population of microorganisms, has emerged as a powerful centerpiece among the methods designed to gain access to the
physiology and genetics of uncultured organisms. Metagenomics has been defined as “the
genomic analysis of microorganisms by direct extraction and cloning of DNA from an
assemblage of microorganisms” (Handelsman, 2004), and its importance stems from the
fact that 99% or more of all microbes are deemed unculturable. Metagenomics promises
to lead to the discovery of new genes that have useful applications in biotechnology and
medicine (Steele and Streit, 2005).

Direct isolation of genomic DNA from an environment circumvents culturing
the organisms under study, and cloning of it into a cultured organism captures it for
study and preservation. Advances have derived from sequence-based and functional
analysis in samples from water (Alio, 2006a) and soil (Xu, 2006). The idea of
cloning DNA directly from environmental samples was first proposed by Pace
(1985), and in 1991, the first such cloning in a phage vector was reported (Schmidt et
al., 1991). The next advance was the construction of a metagenomic library with
DNA derived from a mixture of organisms enriched on dried grasses in the
laboratory. Their application in analyzing unknown environmental DNA samples
have opened a floodgate of exciting research findings (Oremland et al., 2005; Maron
et al., 2007). The phylogenetic analysis of environmental microbial diversity was an
eyear form of metagenomics. Over the years, several significant trends for
metagenomic studies have emerged (Cowan et al., 2005; Galvao et al., 2005; Sharma
et al., 2005; Delwart, 2007).
2.7.4 Approaches to metagenomic analysis

Metagenomic analysis involves isolating DNA from an environmental sample, cloning the DNA into a suitable vector, transforming the clones into a host bacterium, and screening the resulting transformants (Fig. 2.4). The clones can be screened for phylogenetic markers or “anchors,” such as 16S rRNA and recA, or for other conserved genes by hybridization or multiplex PCR or for expression of specific traits, such as enzyme activity or antibiotic production (Courtois et al., 2003; Diaz-Torres et al., 2003; Knietsch et al., 2003; Schloss and Handelsman, 2003), or they can be sequenced randomly (Tyson et al., 2004; Venter et al., 2004). Each approach has strengths and limitations; together these approaches have enriched our understanding of the uncultured world, providing insight into groups of prokaryotes that are otherwise entirely unknown.

Fig. 2.4: Construction and screening of metagenomic libraries. Schematic representation of construction of libraries from environmental samples. The images at the top from left to right show bacterial mats at Yellowstone, soil from a boreal forest in Alaska, cabbage white butterfly larvae, and a tube worm (Handelsman, 2004).
It can be expected that the number of novel genes identified through metagenome technologies will exceed the number of genes identified through sequencing individual microbes (Fig. 2.5).

Fig. 2.5: Timescale of metagenomic-derived and published DNA sequences. The timescale ranges from 1991, the initial outline of the major working steps (Kruger et al., 2003), to the first mapping of archaeal comids in 2002 (Quaiser et al., 2002) and the snap shot sequence analysis of the Sargasso Sea published in 2004 (Venter et al., 2004) (Streit and Schmitz, 2004).

2.7.5 Challenges for metagenomics

Metagenomics is a burgeoning field with new challenges encountered at every instance. The gamut of challenges runs from inefficiencies in sampling, DNA extraction methods, and construction of libraries to inadequacies in data analysis and visualization tools. Added to this are severe computational power and data storage constraints due to the huge amounts of genomic data flooding in from initiatives worldwide. Few of the challenges are described below.
Phylogenetic anchors

The ideal phylogenetic anchor would be equally represented in all species. The 16S rRNA genes do not meet this standard because microorganisms differ in the number of *rrn* operons they carry in their genomes, with a range of 1 to 15. If the number of *rrn* operons is positively correlated with growth rate, as has been postulated (Klappenbach et al., 2000), then slow-growing, difficult to culture bacteria would be poorly represented in 16S rRNA libraries generated by PCR, and their 16S rRNA genes would occur less often in metagenomic clones than the 16S rRNA genes of their rapidly growing counterparts. Once genomes have been reconstructed, one *rrn* operon per genome is sufficient to determine the phylogenetic affiliation of the source of the genes in the genome, but in the absence of metagenome reconstruction for an entire environment, phylogenetic anchors that are found at frequent intervals in genomes are essential. When it is not possible to identify 16S rRNA gene, another anchor is needed. An alternative to finding phylogenetic anchors on a DNA fragment encoding a function of interest is to find fragments of the genome that are linked to the one of interest and search for phylogenetic anchors on them. This method increases the effective size of the contiguous piece of DNA that is being analyzed without requiring an increase in the size of the inserts in the library. To find fragments that are linked in a simple community, a library with high redundancy is needed. The clones can be blotted on a membrane and then probed with the clone of interest to identify those that have an end overlapping with it or clones can be screened by PCR for overlapping regions (Quaiser et al., 2002).

Size of metagenomes

Constructing metagenomic libraries from environmental samples is conceptually simple but technically challenging. If seawater contains 200 species per ml (Curtis et al., 2002), then the metagenome would contain 1 Gbp of unique DNA. To obtain greater than single sequencing coverage, the size of a metagenomic library would need to be many times the size of the metagenome. Because members of a community are not equally represented, it is likely that a metagenomic library of minimum coverage would only represent the genomes of the most abundant species. To obtain substantial representation of rare members (<1%) of the community, the library would likely need to contain 100 to
1000-fold coverage of the metagenome. A library of 500 Gbp might be required to capture the species richness in 1 ml of seawater. Cloning the metagenome of soil, with a species richness 20-fold higher than seawater (Curtis et al., 2002), would be a considerably more daunting prospect (10,000 Gbp). These examples illustrate several challenges in constructing and interpreting information from metagenomic libraries: (a) a large amount of DNA must be isolated and cloned from a small sample, (b) many clones and sequences must be processed to provide meaningful data, and (c) log normal-type population distributions make it difficult to represent the minor species from a sample. Each of these challenges is being addressed, and several studies evaluating methods for library construction have been published (Berry et al., 2003; Gabor et al., 2003; Handelsman et al., 2003).

**Size of inserts**

Strategies for library construction vary depending on the intended study of the resulting library. Libraries containing large DNA fragments are constructed in lambda phage, cosmid, fosmid, and BAC vectors. Small-insert libraries have a significant advantage over large insert libraries because to obtain small inserts, the microorganisms can be lysed by harsh methods that would shear DNA too much to obtain large inserts. Bead beating, for example, extracts DNA from diverse Bacteria and Archaea, providing a good representation of the community. Small-insert libraries are not useful for capturing complex pathways requiring many genes, but they provide an appropriate resource for discovery of new metabolic functions encoded by single genes and for reconstructing metagenomes (Tyson et al., 2004; Venter et al., 2004). The two metagenome reconstruction studies illustrate the difficulty in representing all of the members of a community in the library. As with any genomics study, gene product toxicity is a concern in metagenomic analyses. However, a high-copy vector was used to construct the AMD metagenomic library (Tyson et al., 2004), and the small number of gaps suggests little impact on the composition of the library. High-copy vectors are useful because it is easier to obtain sufficient DNA for further analysis, and for this reason, several plasmids that have copy number that can be modulated from 1 to 50 per cell have been developed.
(Handelsman et al., 2003). These plasmids can be maintained at low copy during clone isolation to avoid toxicity and then can be amplified for screening or plasmid isolation.

**Identifying sequences of interest in large metagenomic libraries**

The library size required to obtain sufficient coverage of the metagenome of even the simplest community presents a significant challenge for screening. Brute-force sequencing (Tyson et al., 2004; Venter et al., 2004) has provided tremendous insight into the libraries and the communities from which they were derived, but the information gleaned from sequencing is limited by the annotation of genes in the existing databases and the available sequencing capacity. Functional screening has the potential to identify interesting genes that would not be recognizable based on their sequences, but sequence-based screening can identify sequences that would not be expressed in the host species carrying the library. A combination of sequence-based methods and functional screening is critical to advancing the field because neither can define the full diversity of gene function in the libraries. High-throughput methods are needed to identify clones carrying functionally active genes, phylogenetic anchors, and novel genes.

**Functional screening:** Advances in screening for active clones will increase the knowledge mined from metagenomic libraries. Key approaches will include new selections that facilitate identification of active clones from among millions of clones. Another productive approach is to construct reporter fusions that respond to expression of the genes of interest. If clones expressing the reporter can be identified rapidly by selection for antibiotic resistance or fluorescence-activated cell sorting, then libraries of sufficient size to represent the diversity of a natural environment could be screened. Implementation of such screens will provide comprehensive functional information that will complement complete sequencing of the metagenome.

**Sequence-based screening:** Screening libraries for genes of interest using primers or probes based on conserved sequences identifies homologues of known genes. This has proved effective to identify phylogenetic anchors and genes encoding enzymes with highly conserved domains (Piel, 2002). A challenge associated with screening libraries for clones carrying phylogenetic anchors is detecting the anchor on the cloned DNA without detecting the homologue in the chromosome of the host cell. This can be
circumvented by using a vector with high or inducible copy number (Handelsman et al., 2003). When the cloned gene is in 50-fold excess to the chromosomal copy, the signal is sufficient to detect.

Approaches are directed toward identifying sequences that are unique to uncultured microorganisms or those specific to a particular environment. These methods involve profiling clones with microarrays that identify previously unknown genes in environmental samples (Sebat et al., 2003), subtractive hybridization to eliminate all sequences that hybridize with another environment, or subtractive hybridization to identify differentially expressed genes (Galbraith et al., 2004), and genomic sequence tags (Dunn et al., 2002). These methods will enhance the efficiency of screening and aid in identifying minor components in communities and genes that define community uniqueness.

### 2.7.6 Major breakthroughs of metagenomics studies

**Global ocean sampling (GOS) expedition**

The world’s oceans contain a complex mixture of micro-organisms that are for the most part, uncharacterized both genetically and biochemically. Rusch et al., 2007 reported a metagenomic study of the marine planktonic microbiota in which surface (mostly marine) water samples were analyzed as part of the Sorcerer II Global Ocean Sampling expedition. A total of 41 samples were collected across 8,000 km transect from the North Atlantic through the Panama Canal and ending in the South Pacific yielded an extensive dataset consisting of 7.7 million sequencing reads (6.3 billion bp). Though a few major microbial clades dominate the planktonic marine niche, the dataset contains great diversity with 85% of the assembled sequence and 57% of the unassembled data being unique at a 98% sequence identity cutoff. Using the metadata associated with each sample and sequencing library, they developed new comparative genomic and assembly methods. One comparative genomic method, termed “fragment recruitment,” addressed questions of genome structure, evolution, and taxonomic or phylogenetic diversity, as well as the biochemical diversity of genes and gene families. A second method, termed “extreme assembly” made possible the assembly and reconstruction of large segments of abundant but clearly nonclonal organisms. Within all abundant populations analyzed,
they found extensive intra-ribotype diversity in several forms: (1) extensive sequence variation within orthologous regions throughout a given genome; despite coverage of individual ribotypes approaching 500-fold, most individual sequencing reads are unique; (2) numerous changes in gene content some with direct adaptive implications; and (3) hypervariable genomic islands that were too variable to assemble. The intra-ribotype diversity was organized into genetically isolated populations that have overlapping but independent distributions, implying distinct environmental preference. They presented novel methods for measuring the genomic similarity between metagenomic samples and showed how they may be grouped into several community types (Rusch et al., 2007).

**Colony collapse disorder**

A study done by Cox-Foster et al., 2007, published in *Science* used metagenomics to probe the aetiology of colony collapse disorder (CCD) in honeybees, which is a widespread economically relevant problem for beekeepers. A dramatic decrease in the number of adult bees occurs in a CCD hive. Four geographically distant hives with CCD were analysed together with a healthy bee hive and royal jelly samples. Unbiased high-throughput pyrosequencing technology was used to produce an inventory of the microflora that was present in CCD and non-CCD hives. A strong correlation was found between the presence of the RNA virus Israeli acute paralysis virus (IAPV) and CCD. This study provides a basis for further investigation into the role of IAPV in CCD and might also establish a strategy for investigating epidemics of unexplained infectious disease (Cox-Foster et al., 2007).

**Sargasso sea microbiome**

One of the most extensive microbial metagenomic studies in the ocean was the shotgun sequencing of microorganisms of size ranges from 0.1 to 3.0 μm in the Sargasso Sea in the Atlantic Ocean near Bermuda (Venter et al., 2004; Tringe et al., 2005; Foerstner et al., 2006). Marine organisms represent the closest living descendants of the progenitor forms of life and contribute to key elementary budgets in the biogeochemical cycles of our planet. Sargasso Sea was also a well studied marine environment and predicted to possess “relatively” low microbial species diversity due to nutrient-poor
conditions. A random sequencing approach was used to generate more than 1625 Mbp of data (almost 1000 times more than the first sequenced bacterial genome in 1995, 1.8 Mbp from *Haemophilus influenzae*) from more than 1,800 different bacterial species (including 148 novel bacterial phylotypes) and encoding over 1.2 million new gene sequences. One of the highlights of the study was the expansion of a family of proteorhodopsin genes. Seven hundred and eighty two (782) rhodopsin like photoreceptors were identified. The discovery of an expanded set of proteorhodopsins with different spectral properties has potential biotechnological applications in optical data storage and signal processing. The data also revealed the presence of an ammonium monooxygenase gene (for ammonium oxidation, producing nitrite) in archaea-associated assemblies; causing reevaluation of the precept that oceanic nitrification is an exclusively bacterial occupation. The catalog of genes generated from this study build an all-essential framework for subsequent genomic comparisons by enriching gene databases with information from unexplored organisms. They also provide a surplus of novel genes or molecular structures that have potential in the development of novel biocatalysts (Venter et al., 2004).

**Acid mine drainage biofilm**

Acid mine drainage (AMD) is a worldwide ecological disaster resulting from commercial mining operations and exacerbated by the activity of uncultured extremophiles that flourish in this inhospitable environment. Acid mine drainages are semi natural environments rich in extremophiles (Galperin, 2004). The metagenomic analyses of a single biofilm sample from an acid mine drainage from the Richmond Mine at Iron Mountain, California, have provided important insights into the microbial community structure (Tyson et al., 2004; Foerstner et al., 2006). A random sequencing approach was used to sequence the metagenome of a microbial biofilm community from AMD in Iron Mountain, California. In striking contrast to the Sargasso Sea survey, the AMD biofilm consisted of only five dominant species, and 75 Mbp of data was sufficient to reconstruct two near-complete genome sequences and gather detailed information about metabolic pathways and even strain-level variation. These included the dominant bacterium *Leptospirillum* group II and the dominant archaeon, *Ferroplasma*
acidarmanus. Ferroplasma is a group of cell wall-less prokaryotes. These two species were also found to be dominant in this community by other analytical methods. In addition to the above two genomes, other reconstructed partial genomes were also identified, including that of a group III Leptospirillum, and an unknown species in the genus Sulfobacillus (Tyson et al., 2004).

**Methane oxidizing archaea from deep sea sediments**

Anaerobic methane oxidation by archaea in marine deep sediments plays a major role in reducing methane (a greenhouse gas) released from oceans into the atmosphere. A metagenomic study of such uncultured communities from a methane seep in Eel River Basin off the California coastline was undertaken. This study utilized an enrichment step to reduce complexity of the sample and select for archaeal DNA. The enrichment step involved size selection (for archaeal cells and sulfate-reducing bacteria) by density centrifugation and size-fractionation prior to clone library construction. Random and targeted sequencing of fosmid clones produced ~ 120 Mbp of DNA that was examined for methanogenesis pathways. An overwhelming absence of a single key enzyme of the 7-step methanogenesis pathway in these samples lent support to a previously proposed “reverse-methanogenesis” hypothesis (Hallam et al., 2003; 2004). Conjecture is that a forward methanogenesis pathway may have been altered or reversed to give rise to the remarkable anaerobic methane oxidation capabilities of methanotrophs (Hallam et al., 2004).

**Metagenomics of the deep Mediterranean ocean**

Most marine metagenomic analyses have been nearly exclusively devoted to photic waters. Martín-Cuadrado et al., 2007 constructed a metagenomic fosmid library from 3,000 m-deep Mediterranean plankton, which is much warmer (~14°C) than waters of similar depth in open oceans (~2°C). They analyzed the library both by phylogenetic screening based on 16S rRNA gene amplification from clone pools and by sequencing both insert extremities of ca. 5,000 fosmids. Genome recruitment strategies showed that the majority of high scoring pairs corresponded to genomes from Rhizobiales within the Alphaproteobacteria, Cenarchaeum symbiosum, Planctomycetes, Acidobacteria,
Chloroflexi and Gammaproteobacteria. They have found a community structure similar to that found in the aphotic zone of the Pacific. However, the similarities were significantly higher to the mesopelagic (500–700 m deep) in the Pacific than to the single 4000 m deep sample studied at this location. Metabolic genes were mostly related to catabolism, transport and degradation of complex organic molecules, in agreement with a prevalent heterotrophic lifestyle for deep-sea microbes. However, a high percentage of genes encoding dehydrogenases and, among them, cox genes, suggesting that aerobic carbon monoxide oxidation may be important in the deep ocean as an additional energy source were observed. The comparison of metagenomic libraries from the deep Mediterranean and the Pacific ALOHA water column showed that bathypelagic Mediterranean communities resemble more mesopelagic communities in the Pacific, and suggests that, in the absence of light, temperature is a major stratifying factor in the oceanic water column, overriding pressure at least over 4000 m deep. Several chemolithotrophic metabolic pathways could supplement organic matter degradation in this most depleted habitat (Martín-Cuadrado et al., 2007).

**Human distal gut communities**

The adult human intestine contains trillions of bacteria, representing hundreds of species and thousands of subspecies. Little is known about the selective pressures that have shaped and are shaping this community's component species, which are dominated by members of the Bacteroidetes and Firmicutes divisions (Xu et al., 2007). The human gut hosts a large community of microbes that are an integral part of human physiology. These organisms may encode >100 times as many genes as the human genome itself. Random sequencing of DNA libraries created from fecal flora of two healthy humans generated ~78 Mbp of data. A comparison of this metagenomic data with that of the human genome and sequenced prokaryotic genomes revealed a relative abundance of pathways for methanogenesis, vitamin synthesis, and degradation of polysaccharides and xenobiotic compounds (e.g., plant glycans obtained through diet, chlorinated organic toxins). These findings have implications for human susceptibility to cancer, obesity, drug metabolism, etc. (Gill et al., 2006).
**Symbiont community from marine worm**

Like humans, lower eukaryotes also enjoy symbiotic relationships with various prokaryotic microbial species. The metagenome of bacterial symbionts of a marine oligochaete worm (*Olavius algarvensis*) that has uniquely evolved to do without a mouth, gut, or excretory system, was sequenced. The hypothesis is that these bacterial symbionts compensate for the loss of digestive and excretory systems in their host. ~204 Mbp of randomly sequenced data from small and large clone libraries allowed partial assembly of four symbiotic species. Reconstruction of their physiologies from the genomic data revealed extensive carbon fixing capabilities amongst others that may fulfill their host’s energy and waste management requirements (Woyke et al., 2006).

### 2.8 Novel biotechnological and pharmaceutical products

The information held within a metagenomic library can be used to determine community diversity and activity, the presence of specific microorganisms or biosynthetic pathways as well as simply searching for the presence of individual genes (Fig. 2.6).

![Fig: 2.6: Global metagenome searches will achieve two major goals.](image)

Firstly, they will result in a tremendous increase of knowledge on the function of microbial niches. It is probably the only possible way to link the natural biodiversity with functions of non-cultured microbes in the respective niches and explore their roles in the environmental processes. Secondly, metagenome research will lead to the discovery of many novel enzymes with new functions. Further, it can be expected that metagenome research will identify novel drug molecules and other valuable metabolites (Steele and Streit, 2005).
Because of the overwhelming majority of non-cultured microbes in most microbial niches, metagenome searches have resulted in identification of unknown genes and proteins. The probability of uncovering novel sequences makes this approach more favorable than searches in cultivated microbes. Genes involved in essential pathways (such as nitrogen, carbon dioxide fixation and iron metabolism) were revealed. The genomic sequence data identified genetic polymorphisms for many genes and suggested evidence for genetic recombination in the microbial populations. The metagenome sequence information established a foundation for comparisons of microbial communities. In addition, a recent proteomic analysis identified an abundant novel protein, a cytochrome, as an essential component to iron oxidation and acid mine drainage formation (Ram et al., 2005). These results have the potential to guide the remediation of sites contaminated by acids. Metagenomics has identified a significant number of novel genes encoding for biocatalysts or molecules with high potential for use in pharmaceutical products or production processes. The genes of interest within these searches are often the key genes involved in the synthesis of antibiotics by Gram-positive Streptomyces and are often part of large biosynthetic gene clusters. Lipases, esterases, polysaccharide-modifying enzymes such as the starch modifying enzymes, isolation of enzymes useful for the production of bulk proteases are of considerable interest for biotechnological applications (Jaeger and Eggert, 2002; Lammle et al., 2006; Hardeman and Sjoling, 2007).

**Phylogenetic analyses of environmental ribosomal RNA**

The use of culture-independent methods to estimate microbial diversity in the environment started in the 1980s (Pace et al., 1985). This involved isolating total DNA directly from the environment, cloning the DNA using vectors and screening for clones that hybridized to the rRNA probes, and sequencing the positive clones. Many types of rRNA sequences not present among cultured microbes from the same samples were identified. The incorporation of gene-specific PCR before the cloning step significantly streamlined the procedure and allowed more direct estimation. The very first application of PCR in phylogenetic analysis of mixed microbial communities in ocean waters led to the discovery of ubiquitous and abundant groups of new micro-organisms (Giovannoni et
al., 1990). This study identified significant genetic heterogeneity among closely related phylogenetic types. Phylogenetic comparisons of rRNA genes from environmental sources have led to the discovery of many novel microbial taxonomic groups. In 1987, based on rRNA sequence data, Woese identified 12 major divisions (phyla) in the domain bacteria. The analysed bacteria represent almost all major cultured groups of bacteria for microbiological research. In over a decade, culture-independent surveys identified that there are at least 40 well major bacterial divisions. There are about 30 bacterial divisions with no or very few cultured representatives (Hugenholtz et al., 1998; Konstantinidis et al., 2006). These discoveries are now guiding a coordinated effort by the microbiology community to culture representatives from many of the unknown major divisions of bacteria in order to study their genetic, physiological and ecological properties.

The culture-independent methods have also revealed major new types of archaea. At present, there are about 300 cultured archaeal species, primarily belonging to phylum euryarchaeota. Schleper et al., 2005 compiled over 8000 deposited archaeal rRNA gene sequences from various natural environments. Phylogenetic analyses suggested that domain archaea contains at least 50 distinct phylogenetic groups. The divergence among these phylogenetic groups is similar to those among many bacterial phyla. Among these 50 phylogenetic groups, only 13 have cultured representatives. Before the application of culture-independent methods, archaea are thought to be only present in extreme habitats. Recent investigations have identified that archaea are also widespread in diverse non extreme habitats; such as gardens and forests, water and sediments in marine and freshwater lakes, as well as extreme habitats such as hot springs, saline lakes and deep-ocean thermal vents (Black Smokers).

In the marine environment at depths 100–5000 m, the average microbial density accounts for about 20% of all microbial cells in the ocean. Phylogenetic analysis indicates that bacteria differ significantly from one another on the basis of their rRNA sequences (Huber et al., 2002).

2.10 Richness estimations using DOTUR

Modern molecular techniques have revealed an extraordinary diversity of microorganisms, a major challenge to microbial ecologists is how to compare the
microbial diversity of different environments when the vast majority of microbial taxa are usually unknown? The statistical approaches developed by ecologists and evolutionary biologists are proving to be promising tools to meet this challenge. The combination of these tools with molecular biology techniques allow the rigorous estimation and comparison of microbial diversity in different environments.

Several approaches to compare diversity have been applied recently to molecular studies of microbial diversity. Three of these newly rediscovered approaches are: parametric estimation, nonparametric estimation and community phylogenetics. The first two approaches are used to compare OTU richness among environments. The third approach compares evolutionary diversity among environments. Each approach has its own unique strengths and limitations. We will discuss only parametric methods, which we have used to analyze our data from isolates and clones from various environmental samplings.

**Parametric methods**

Parametric approaches estimate the number of unobserved OTUs in a community by fitting sample data to relative OTU abundance (species abundance). The advantage of this approach is that, we can use this method to estimate diversity from relatively small samples of individuals from a given environment. Thus, this approach is ideal for estimating the diversity of hyperdiverse organisms such as microbes (Bohannan and Hughes, 2003).

**Rarefaction curves**

Nucleotide sequences provide a more precise analysis and these sequences are usually grouped as OTUs or phylotypes. Sequences with greater than 97% identity are typically assigned to the same species, those with less then 95% identity are typically assigned to the same genus, and those with less then 80% identity are assigned to the same phylum (Hugenholtz, 1998; Sait et al., 2002) A genetic distance is approximately equal to the converse of the identity percentage. These cutoff values are a best fit of taxonomy with modern 16S rRNA gene sequencing. There are few methods available to assign sequences to OTUs quickly based on sequence data (Seguritan and Rohwer, 2001).
One method of comparing 16S rRNA sequences is to calculate distances between known and unknown sequences. Distance values of 0.03 are thought to differentiate at the species level, 0.05 at the genus level, 0.10 at the family/class level, and 0.20 at the phylum level. A species is described as a group of sequences that are within a distance of 0.03 of each other.

To assign sequences quickly and accurately to OTUs, a computer program DOTUR (Distance-Based OTU and Richness) is available (http://www.plantpath.wisc.edu/fac/joh/dotur.html).

Distance matrices by using DNADIST from the PHYLIP package with the Jukes-Cantor correction for multiple substitutions (http://evolution.genetics.washington.edu/phylip.html) were constructed. A computer program, DOTUR (Distance based OTU and Richness) was developed that uses a furthest-neighbor (complete-linkage) algorithm to assign sequences into OTUs and then constructs rarefaction curves for each distance level (http://www.plantpath.wisc.edu/fac/joh/dotur.html) (Schloss and Handelsman 2005a). Distance matrices from DNADIST are used as input files for DOTUR. It then calculates values that are used to construct randomized rarefaction and collector’s curves of observed OTUs, diversity indices, and richness estimators. In my work, I have used DOTUR by analyzing and comparing 16S rRNA gene identified from various environmental isolates and library, which have not been studied previously. In clone library and individual isolates, DOTUR assigned sequences to OTUs more accurately and consistently (Schloss and Handelsman, 2005a).

**Interpreting the richness**

Construction of rarefaction curves for each phylum compares the extent of sampling of each phylum at various taxonomic levels and their relative richness. For example, rarefaction curves from OP11, which currently has no cultured representatives (Fry, 2000), Acidobacterium, which is one of the most abundant phyla in soil but is difficult to culture, and γ-Proteobacteria, which is the most, well-sampled and well-studied phylum, indicates that the rate of discovering new sequences remains high for all phyla considered. It is possible to use rarefaction to determine differences in relative
richness between phyla. OP11, has been shown to have a patchy distribution in various environments, has relative species richness higher than that of the Acidobacteria but lower than that of the γ-Proteobacteria (Schloss and Handelsman, 2005b).

2.11 The Shannon-Weiner diversity index

**Species diversity (H):** The number of different species in a particular sample (i.e. species richness) weighted by some measure of abundance such as number of individuals or biomass.

**Species richness (S):** The total number of different species in a particular sample.

**Species evenness (E):** The relative abundance with which each species are represented in a sample.

\[ H = \sum (pi) (\log_2 p - i), \]

where \( p \) represents the proportion of a distinct phylotype relative to the sum of all distinct phylotypes.

Evenness (E) was calculated as: \( E = H/H_{max} \), where \( H_{max} = \log_2 (S) \).

Richness (S): Total number of species in the community.

Species richness and species evenness are probably the most frequently used measures of the total biodiversity of a region. An ecosystem where all the species are represented by the same number of individuals has high species evenness. An ecosystem where some species are represented by many individuals, and other species are represented by very few individuals has a low species evenness. For example, some areas may be rich in closely related taxa, having evolved from a common ancestor that was also found in that same area, whereas other areas may have an array of less closely related species descended from different ancestors (Magurran, 1998).