METAGENOMICS FOR ENVIRONMENTAL AND INDUSTRIAL MICROBIOLOGY

RINKOO D. GUPTA AND RAKESH SHARMA*

Our environment is a vast reservoir of a large number of microorganisms and only a minute fraction of them can be cultured. Metagenomics is a culture-independent genomic and functional analysis of this unculturable microbiota. Metagenomics, being a pool of genomes of highly diverse unculturable microorganisms, has increased the probability to discover genes and pathways for several new enzymes with high and specific catalytic properties, bioactive compounds and bioremediation processes. Metagenomic studies are mainly based on three types of analyses- a function-based screening of metagenomic libraries for an expressed trait, a sequence-based screening of metagenomic libraries for particular DNA sequences and large-scale sequencing and in-silico analysis of the data for particular traits. This review aims to highlight the recently developed metagenomic approaches to the environmental and industrial microbiology.

Introductory

Our world contains a huge diversity of unculturable microbiota which provide a massive repertoire of novel enzymes and biomolecules that can help in the development of environmental biotechnology and industrial growth. Metagenomics has emerged as a new tool with an enormous potential in environmental and industrial biotechnology research. ‘Metagenomics’ is a combination of two words- Meta, which is a process of statistically combining separate analyses and Genomics, a comprehensive analysis of an organism’s genetic material1, 2. Metagenomics differs from the traditional genome analyses where a single genome is analyzed at a time. To explore the diversity of a microbial community, mainly 16S rRNA sequence was used as a marker due to the presence of often conserved sequences within a species. 16S rRNA sequencing revealed that there are large groups of microorganisms in many environments which are yet to be cultured and studied3. Recently described, a gene-centric approach considers a community from the viewpoint of its component genes rather than considering a community from the point of view of its component organisms (genomes). Genes that are found more frequently in one community than another are assumed to endow a beneficial function on that community3. This kind of gene-centric approach can augment the chances of mining novel enzymes and bioactive molecules from these microorganisms.

Currently, the process of metagenomics includes the extraction of genomic DNA directly from the environmental samples where these microorganisms are present and their analyses on the basis of functions and sequences4,5. The metagenomic DNA can be used to prepare metagenomic libraries in plasmid, cosmid or bacterial artificial chromosome based vector. Function-based metagenomics relies on cloning environmental DNA into expression vectors and propagating them in appropriate hosts, followed by appropriate activity screens. Activity based screenings of metagenomic libraries are mainly the cultivation of metagenomic clones on indicator plates which allows the analysis of defined enzyme activities via biocatalytic conversion of an indicator substrate that leads to the formation of a clear or colored area surrounding the
active colony. In principle, activity-based screening of metagenomes requires the concerted expression of all environmental genes located on a given DNA fragment, irrespective of its size\(^6,7\). This approach does not require the sequence of the gene and has potential to identify entirely new classes of genes for new or known functions. In spite of the advantages, functional approach has several limitations: most of the genes isolated directly from environmental samples cannot be expressed well and it is difficult to find out a suitable screening method. Craig et al. has expanded small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria\(^8\).

Sequence based metagenomics requires high throughput sequencing using traditional Sanger’s method or next generation sequencers (Roche FLX, Solexa or Solid) of the metagenomic DNA. Next generation sequencing methods do not require cloning of the DNA before sequencing and generate far more sequence data per run, at a much lower cost than conventional sequencing, but the reads are shorter\(^3\). The improvements in these technologies are producing longer reads, and single-molecule sequencing holds the promise of even longer reads.

**Environmental Applications**

Several years of long industrial activity has resulted to the emergence of new pollutants in the environment. These artificially synthesized toxic compounds are a challenge to the capacity of the microbial communities of the contaminated sites. There are several examples where the release of these compounds into the environment has produced unpredicted adverse effects. Eco-friendly and cost-effective degradation and/or remediation of these pollutants are being explored by exploiting the biodegradation property of microorganisms.

Biodegradation is the ability of microorganisms to remove complex chemicals from the environment through a complicated process in which many biotic and abiotic factors are implicated into the natural biogeochemical cycles. Biodegradation converts complex organic compounds into simpler ones with the help of microorganisms and thus removing the complex chemical compounds from the environment. Microbial bioremediation is a well-known strategy for the degradation of anthropogenic compounds from the polluted environments. Metagenomics is being used as a mode to explore unculturable bacteria from the contaminated soil and water bodies which are adapted for the contaminated sites by acquiring the capacity to degrade the toxic chemicals. Identification of the genes involved in the degradation of toxic chemicals and the discovery of new metabolic pathways for the complete mineralization of the pollutants would be of huge importance for bioremediation and industrial application. However, classical remediation techniques (land filling, incineration, etc.) are generally expensive and not always efficient and bioremediation of the industrial contaminants requires a thorough understanding of factors governing the growth, metabolism, dynamics and functions of indigenous microbial communities at contaminated sites.

There are many examples in the literature where metagenomics has helped in the process of degradation of toxic industrial pollutants. Polychlorinated biphenyls, a class of synthetic aromatic compounds, are widely used in the electronics industry and as components of adhesives and plastic materials. These compounds are thermally and chemically stable and found as pollutants throughout the ecosystem and food chain due to their lipophilic behavior, and constitute a serious health problem to mammals. From the polychlorinated-biphenyl contaminated soil, bacteria and functional genes were identified which was associated with biphenyl degradation in the root zone of an Austrian pine by using stable isotope probing\(^9\). Another major contaminant is organophosphorus insecticide, which is widely used for crop protection. The primary degradation product of insecticide, chlorpyrifos, is 3,5,6-trichloro-2-pyridinol (TCP). Because TCP is more polar and water soluble than chlorpyrifos, it is more mobile in soil and more leachable into groundwater and surface water and hence more hazardous to human and animal health. Recently metagenomics approach was applied to isolate a novel amidohydrolase gene, tep3A encoding a TCP degrading enzyme from a cow rumen metagenomic library\(^10\). Pyrethroids and pyrethrins are widely used insecticides. Frequent use of these insecticides resulted in resistance in insects. This has also resulted in environmental issues and human exposure. Numerous studies have shown that very high exposure to pyrethroids might cause potential problems to man and aquatic organisms. Therefore, it is important to develop a rapid and efficient disposal process to eliminate or minimize contamination of surface water, groundwater and agricultural products by pyrethroid insecticides. A novel pyrethroid-hydrolyzing esterase gene was successfully isolated from the metagenomic DNA combined with activity-based functional screening from contaminated soil\(^11\). The broader substrate specificities and higher activity of the metagenomic pyrethroid-hydrolyzing esterase make it an ideal candidate for in situ detoxification of pyrethroids.
**Industrial Applications**

Currently, there is a global drive to promote industrial biotechnology as a central feature of the sustainable economic future of modern industrialized societies. Many natural products and their derivatives are produced from bacterial and fungal secondary metabolites. Consequently, metagenomic DNA clone library offers possibilities to discover novel bio-molecules through expression of genes from uncultivated bacteria. Enormous resources of the uncultivated microbial diversity have been exploited for the production of bioactive molecules. Several novel small molecules, antibiotics and new antibacterial proteins have been identified using metagenomic approaches. The recent application of metagenomics to the discovery of bioactive small molecules, small molecule biosynthetic gene clusters and antibacterial active proteins has been discussed in the recent review by Banik and Brady.

Marine sponges are rich source of bioactive natural products with clinical potential. They harbour rich communities of symbiotic bacteria as true producers of sponge-derived compounds. Polyketide synthase, one of the most important groups of sponge-derived drug biosynthesis genes, were identified from different sponge metagenomes. Metagenomics has an invaluable role in the discovery process of new bioactive molecules and promises to provide new molecules with diverse functions, but ultimately, proper expression systems are required for economic success. In future, a successful attempt to metabolically engineer *E. coli* to improve its use as a cloning and production host for complex secondary metabolites is quite probable.

Metagenomics has increased the probability to discover several novel enzymes with high and specific catalytic properties sought for various industrial applications like synthetic organic chemistry, chiral resolutions, food and flavour, biofuels and synthetic biology. Plenty of novel genes encoding industrially important enzymes have been isolated from the metagenome. Recent reviews summarize the discovery of biocatalysts using metagenomics.

A cost effective production of second-generation biofuels has been a hot spot of research and industries since a decade. Biofuels have attracted great interest as an alternative, renewable source of energy. Energy dependence on ongoing fossil fuels and growing environmental awareness of the critical consequences of burning such fuels has forced the necessity of production of cost effective, eco-friendly and renewable resources. Plant biomass, which is the most abundant biopolymer containing the largest reservoirs of organic carbon on Earth is largely inaccessible to most organisms. These biopolymers, mainly present in the form of celluloses, hemicelluloses, and lignin, have been well-recognized as a potential sustainable source for biofuel production. Examples of microbial degradation of major plant biomass i.e. lignocelluloses for biofuel production are diverse. Still our current understanding of the enzymes involved in these processes is limited to a handful of model organisms such as the fungus *Trichoderma reesei* and the bacterium *Clostridium thermocellum*. There are several biofuels which have already been produced commercially e.g. bioethanol, biobutanol, biodiesel and biogases; however cost-effective production methods are still elusive. In order to identify novel microbes and enzymes to accomplish the goal of cost-effective production, metagenomic libraries from several resources are being targeted.

For biofuel production, lignocellulose degrading microbial enzymes have been identified from bovine rumen microbiome by applying gene centric metagenomic approach. The bovine rumen and gastrointestinal tract harbour a diverse and complex microbiome which enables to identify specific novel enzymes for forage degradation. The identified genes, glycoside hydrolase and cellulosome functional genes were compared and found to be very specific to the easily available side chains of complex plant polysaccharides and not the more recalcitrant main chains, especially cellulose. These substrate specific microbial enzymes would allow engineering novel pathways for the degradation of specific polysaccharides for biofuel production.

The main obstacle to commercial production of biofuels is to design suitable microbial hosts which can tolerate process stresses such as end-product toxicity and tolerance to fermentation inhibitors in order to achieve high yields and titers. To address this issue, a functional metagenomic approach was applied to screen metagenomic fosmid libraries and identified genes conferring tolerance towards seven important biomass inhibitors. Two different and the most relevant fosmids, showing improved growth up to 7-fold in the presence of inhibitory concentrations of syringaldehyde and 2-furoic acid, respectively, were explored to identify individual genes responsible for the tolerance phenotypes. Out of all the genes identified from the two fosmids, a combination three-genes construct was found to confer tolerance to mixture of these biomass inhibitors. This type of novel genes for inhibitor tolerance can provide a starting point to engineer robust strain for cost-effective and eco-friendly biofuel production.
**Metagenomics Research at IGIB**

We are preparing metagenomic libraries from environmental samples collected from various ecological niches in different geographical regions in India. These libraries are being screened for functional clones expressing novel biocatalysts, bioactive compounds and pathways.

Lipolytic enzymes include lipases and esterases, which hydrolyze short chain acylglycerol and have significant biotechnological importance because of their ability to catalyze regio- and stereo-selective organic reactions. These enzymes are also becoming important in synthetic organic chemistry because of their capability to withstand organic solvents and catalyze reverse reactions. Identification of lipolytic enzymes with a spectrum of sequence and functional diversity works as a suitable biocatalyst for challenging reactions and reaction conditions. Screening of metagenomic libraries representing various ecological niches resulted in isolation of more than 100 lipolytic clones. Many of these lipolytic genes from the metagenomic clones showed low similarity to known and putative lipolytic genes in databases. These genes represent different lipolytic gene families and are expected to show distinct substrate preference. Out of 12 unique lipolytic enzymes with an ability to hydrolyze tributyrin identified from a metagenomic library of pond water microbial assemblage, ten belonged to seven known lipolytic protein families. One of the proteins showed similarity to BioH and another belonged to a yet uncharacterized a/b hydrolase protein family abh_upf0017.

Oxygenases constitute another group of important biocatalysts, involved in varied processes including biodegradation of aromatic compounds and biosynthesis of secondary metabolites. These enzymes are broadly classified as monooxygenases and dioxygenases depending on whether they incorporate one or two oxygen atoms in the substrate. Several dioxygenases and monooxygenases have been reported to catalyze the production of indigo and indirubin. Indigo is one of the most commonly used dyes in the textile and printing industries. Indirubin, a red structural isomer of indigo is an active constituent of traditional Chinese medicine used against chronic myelogenous leukemia and has also been shown to have anti-inflammatory effects. We screened a metagenomic library of effluent treatment plant sludge to find monooxygenases that could be useful for dye production. Two clones encoding proteins with similarity to putative flavin monooxygenases from *Mesorhizobium loti* and *Sphingomonas wittichii*. They represented flavin-containing monooxygenases and Baeyer–Villiger monooxygenases subfamilies. The two proteins have high potential for use in biotransformations and indigo production.

Most bacteria in their natural habitats encounter various abiotic stress conditions. Possession of distinct stress tolerance mechanisms provides them an edge over their competitors to survive under these stress conditions. Metagenomics plays a pivotal role in identification of these mechanisms of stress tolerance and their role in adaptability of the unculturable bacteria in a particular environment. We enriched *Escherichia coli* clones with an ability to grow at inhibitory NaCl concentration from a pond water metagenomic library. From two unique clones, genes encoding for proteins with similarity to putative general stress protein (*GspM*) harbouring GsiB domain and putative enoyl-CoA hydratase (*EchM*) were identified to be responsible for salt tolerance. These genes have potential application in generating salt tolerant recombinant bacteria or transgenic plants.

Currently, the main focus of our research is towards understanding mechanisms of arsenic resistance in unculturable bacteria and to identify their role in arsenic mobilization in contaminated environment. Two arsenic resistant clones were selected from the effluent treatment plant sludge metagenomic library. These clones MT3 and MT6 had 8 and 18 fold higher resistance to sodium arsenate in comparison to the parent strain, respectively. A novel arsenate resistance gene (*arsN*) encoding a protein with similarity to acetyltransferases was identified from clone MT6. **ArsN** homologues were found closely associated with arsenic resistance genes in many bacterial genomes. **ArsN** alone resulted in about 6 fold higher resistance to sodium arsenate in wild type *Escherichia coli* W3110. Recent studies show that microbes may be playing a significant role in mobilization of arsenic from sediment to aquifers. Understanding of microbial diversity, arsenic resistance and genes involved in arsenic mobilization in sediments and aquifers is expected to provide better and may be permanent solutions to hazard posed by arsenic contamination of drinking water.

**Acknowledgements**

We thank Director, IGIB for encouragement and support. Research in RS lab is supported by grants from Council of Scientific and Industrial Research and Department of Biotechnology.
References

1. J. Handelsman, Microbiology and Molecular biology review. 68, 669 (2004)