Universal Journal of Environmental Research and Technology
Available Online at: www.environmentaljournal.org

2011 Vol 1 1-6

Open Access

Review Article

A Mini Review on Metagenomics and its Implications in Ecological and Environmental Biotechnology

[1]Patake R. S. and [2*]Patake G. R

[1]Department of Biotechnology Fergusson College Pune,
[2*]Soham BioinfoTech, Pune, India
*** Corresponding author: kdipali@gmail.com, ravindrapatake@gmail.com**

Abstract

Metagenomics is an applied science that deals with throughput analysis of Environmental Genomic isolates. Advances in genomic sequencing technologies such as parallel sequencing and sequence analysis methods have contributed most in foundation of this field. Metagenomics techniques have broad perspectives in field of Ecological Biotechnology. Its potential applications in in-situ restoration of ecosystems, designing strategies for bioremediation and in monitoring biodetereoration have come to the edge. In this review, we have reviewed current developments in metagenomics methodology with respect to applications in the field of Ecological Biotechnology and Environmental Biotechnologies.

Keywords: High throughput, Metagenome, Metagenomics, Reads

1. Introduction

Micro organism communities are important in the functioning of all ecosystems, but the unculturable microorganisms and their Role in natural ecosystems are unclear (Tyson *et al*., 2004). Microbes obtained from environmental sample in pure culture form are not always representing dominant species (Hugenholtz, 2002). High throughput sequencing of entire array of genomes present in environmental samples enables us to not only enumerate but also classify microorganisms with their phylogenetic relationship among themselves (Mardis, 2008) Metagenomics is based on the genomic analysis of microbial DNA directly from the communities present in samples such as soil, water or faeces. This technology - genomics on a large scale - will probably lead to great advances in medicine, agriculture, energy production and bioremediation. Metagenomics can unlock the massive uncultured microbial diversity present in the environment for new molecules for therapeutic and biotechnological applications. The field of Metagenomics developed as a consequence of the discovery that prokaryotic diversity was much greater than previously realized and that the prokaryotic population was a significant resource for biotechnology and environmental applications, both facts which were reaching the limitations of traditional culture based investigation (Steele and Streit, 2005). There is a growing belief that the term 'unculturable' is inappropriate and that in reality we rather have yet to discover the correct culture conditions. The development of metagenomic technologies over the past five years has provided access to much of the prokaryotic genetic information available in environmental samples, independent of culturability (Cowan, 2005). In this review, we are discussing the methods of Metagenomics which are being practiced in the field of environmental and ecological biotechnology and current applications and development in the same. We will first focus the current technologies and then their implications.

2. Current Technologies

Current technologies in the field can be broadly classified in two categories, wise 1) Sampling and sequencing technologies and 2) Metagenome analysis technologies. For validation of some metagenomics sequence analysis tools, pseudo data generator or simulator algorithms are also being used (Warren, 2007). Sequencing technologies being

used so far are listed in Table 1. Not all but few of them are in the focus of our discussion. Sampling Methods depends on niche and habitat of microorganisms in environment (Abe, 2005). Since 2004, several metagenomics sequencing projects have been successfully implemented, such as Acid Mine Drainage Biofilm (AMD) for dozens of species and the recent Human Gut Microbiome (HGM) for more than thousands of species (Yang *et al*., 2008). The newly reported tiny marine animals that complete their life cycle in the total absence of light and oxygen are members of the phylum Loricifera and they are less than a millimeter in size. Their genomes were collected from a deep basin at the bottom of the Mediterranean Sea, where they inhabit a nearly salt-saturated brine that, because of its density (>1.2 g/cm3), does not mix with the waters above. As a consequence, this environment is completely anoxic and, due to the activity of sulphate reducers, contains sulphide at a concentration of 2.9 mM. Despite such harsh conditions, this anoxic and sulphidic environment is teeming with microbial life, both chemosynthetic prokaryotes that are primary producers, and a broad diversity of eukaryotic heterotrophs at the next trophic level (Mentel and Martin, 2010). Metagenomics was initially employed to study non-culturable microbiota and focused primarily on providing a better understanding of global microbial ecology in different environmental niches. With the advent of efficient cloning vectors such as bacterial artificial chromosomes (BACs) and cosmids, together with improved DNA isolation techniques and advanced screening methodologies using robotic instrumentation; it is now possible to express large fragments of DNA and subsequently screen large clone libraries for functional activities. A clear example such genome collection venture was the large scale metagenome sequencing project which was recently undertaken on oligotrophic seawater samples from the Sargasso Sea and the Global Ocean Sampling (GOS) expedition (Kennedy *et al*., 2010). Central theme of Metagenomic data analysis is presented in Figure 1.

Micro and macro environmental samples are collected and then high-throughput sequencing is done to obtain metagenomic reads. Then all genomes (Metagenomes) are assembled separately using computational algorithms. The array of assembled genome is then subjected to analysis. Following are the some common points which are the end products of metagenome analysis:

1. Phylogenetic relationship among the metagenomes
2. Comparative genomic approaches which includes:
    a. Functional genomics and structural genomics
        i. Metabolomics
        ii. Finding Orthologs
        iii. Finding Epigenetic interaction
3. Niche Mapping on Micro and Marco environment.

Additional innovative screening approaches such as Substrate- Induced Gene Expression screening (SIGEX) have facilitated the cloning of catabolic operons potentially involved in benzoate and catechol degradation among others. These functional based screening approaches have also been supplemented with homology-based screens, primarily involving polymerase chain reaction (PCR)-based approaches targeting novel genes with sequences similar to known genes. This has resulted in the cloning of genes such as polyketide synthases, alkane hydroxylases, cyclomaltodextrinases, xylanases and beta-xylanases. Recently novel methods such as pre-amplification inverse-PCR (PAI-PCR) and metagenomic DNA shuffling have been employed to isolate new biocatalysts. PAI-PCR which has been employed to isolate glycosyl hydrolase genes from horse and termite guts, offers the potential to clone genes for which the copy number of target DNA sequences is low, while the shuffling approach, which has been used to construct novel biocatalysts, simulates and accelerates the evolutionary process using molecular biological tools (Kennedy *et al*., 2010). SNP (Single Nucleotide Polymorphism) analysis can also be done at high throughput scale among the metagenomes and this was reported in the case of pseudomonas metagenome reads (Spencer, *et al*., 2003).

Use of 16s rRNA is also technique of choice when there is use of metagenome reads for phylogenetic analysis. Micro environmental sampling was reported from human fecal microbiota for 16s rRNA analysis (Kurikka, *et al*., 2009). Computational software and web servers are listed in table 2. Effective computational pipeline and methods are being reported. Warren A. et-al has developed methodology based on ORF analysis to find out missing genes in prokaryotic reads annotations (Warren, *et al*., 2010). White J. R. et-al used phylogenetic markers for alignment and clustering to study microbial diversity from environmental samples (White, *et al*., 2010). Molecular paleo-

bacteriology and molecular evolutionary studies are also reported where the metagenomics reads were used (Sun and Caetano-Anolles, 2010). Pyrosequencing- technique that yield rapid metagenome reads, also generate redundant natural and artificial duplicate reads. Discrimination between artificial and natural duplicate or near duplicate reads is computational challenge. Beifang Niu et-al tackled above problem with algorithm that utilizes all against all read comparison and clustering (Niu, *et al*., 2010). Gupta and Mathews used combination of phylogenomic and protein signature based approach to characterize the major clads of cynobacteria (Gupta and Mathews, 2010).

Metabolomic analysis of deep mine microbial ecosystem was successfully done using pyrosequence reads (Edwards *et al*., 2010). Because the sequencing technologies are rapid, the chances of errors in sequence increases in folds when genomes are concerned. The effect of sequencing errors on metagenomic gene prediction was studied by Hoff (2009). Saturated bines are extreme environments of low diversity. *Salinibacter ruber* is the only bacterium that inhabits this environment in significant numbers. In order to establish the extent of genetic diversity in natural populations of this microbe, the genomic sequence of reference strain DSM 13855 was compared to metagenomic fragments recovered from climax saltern crystallizers and obtained with 454 sequencing technology. This kind of analysis reveals the presence of metagenomic islands, i.e. highly variable regions among the different lineages in the population (Pasic *et al*., 2009). Binning of metagenomic fragments was also attempted using the oligonucleotide frequency derived error gradients (Isaam and Saman, 2009).

The first step, which is still a major bottleneck, of metagenomics is the taxonomic characterization of DNA fragments (reads) resulting from sequencing a sample of mixed species. This step is usually referred as “binning”. Existing binning methods are based on supervised or semi-supervised approaches which rely heavily on reference genomes of known microorganisms and phylogenetic marker genes. Due to the limited availability of reference genomes and the bias and instability of marker genes, existing binning methods may not be applicable in many cases. To overcome above bottleneck, Yang Bin et-al used unsupervised binning method based on the distribution of a carefully selected set of l-mers (substrings of length l in DNA fragments) (Yang *et al.,* 2010).

Table 1: Metagenome and other Sequencing Methods

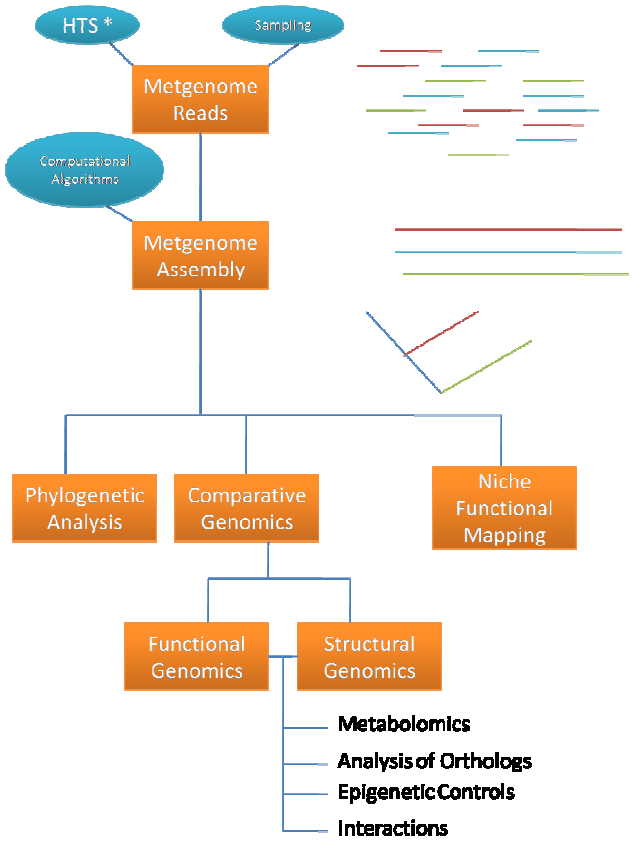
Sequencing Method	Comment	Reference
BAC-based sequencing	Old Method of Genomic era where bacterial Artificial chromosomes were used (scrupulously used between year 1995 to 2002)	Mardis, 2008
WGS (Whole Genome Sequencing) also called as shotgun sequencing.	This technology first time boost the speed of sequencing genome, and was being meticulously used since 1999 to 2006 to sequence whole genome through various project. E.g. Human Genome Project.	Mardis, 2008 and Warren *et al*., 2007
Roche/454/Pyro/FLX Sequencing	Very Rapid parallel sequencing method, being used for metagenomes since 2005 onwards	Mardis, 2008
Illumina/Solexa Genome Analyzer	Very Rapid parallel sequencing method, being used for metagenomes since 2006 onwards, utilizes a sequencing-by-parallel-synthesis approach	Mardis, 2008
Applied Biosystems SOLiD™ Sequencer	Uses an emulsion PCR approach with small magnetic beads to amplify the fragments for sequencing.	Mardis, 2008
Chromatin immunoprecipitation (ChIP) sequencing	Sequences and maps “Reactome” than “Genome”	Mardis, 2008
Helicos Heliscope™ and Pacific Biosciences SMRT	Recently announced massive parallel sequencing and analysis platform	Mardis, 2008

Table 2: Software, Web Based Tools and Databases Used for Metagenome Analysis

Name of Software/web server	Utility	Reference
UniFrac http://bmf.colorado.edu/unifrac/	An online tool for comparing microbial diversity	Lozupone *et al*., 2006
JANE http://jane.bioapps.biozentrum.uni-wuerzburg.de	Mapping of ESTs and variable length prokaryotic genome sequence reads on related templet genomes	Liang *et al*., 2009
WebCARMA http://webcarma.cebitec.unibilefeld.de	A web application for the functional and texonomical classification of unassembled metagenomic reads	Gerlach *et al*., 2009
DraGnET http://www.dragnet.cvm.iastae.edu	Software from sorting, drafting and analyzing annotated draft genome sequence data	Duncan, *et al*., 2010
Prodigal http://compbio.ornl.gov/prodial/	Prokaryotic gene recognition and translation initiation site identification	Hyatt, *et al*., 2010
MEGAN *www.ab.informatik.uni-tuebingen.de/software/megan*	Illumina sequencing metagenome reads analysis tool	Mitra, *et al*., 2010
MG-RAST http://metagenomics.anl.gov/	Metagenome annotation server	Aziz, 2010
CAMERA http://camera.calit2.net	Metagenomic database server which contains sequences from environmental samples collected during the global ocean sampling (GOS)	Maumus *et al*., 2009
FUNGIpath http://www.fungipath.upsud.fr	Database and tool server for fungal orthology and metagenomics	Grossetete *et al*., 2010
envDB http://metagenomics.uv.es/envDB/	Database and tool server for environmental distribution of prokaryotic taxa	Tamames *et al*., 2010

3. Concluding Remarks

Metagenomics is an applied science that deals with the throughput genomic analysis of environmental isolates. Advances in genomic sequencing technologies, such as parallel sequencing and sequence analysis methods have contributed to the very foundation of this field. Metagenomics techniques have broad prospects in the field of environmental biotechnology. Its potential use in the on-site restoration of ecosystems, and developing strategies for bioremediation and monitoring biodetereoration came to the edge. Tools, software and databases are increasing in considerable numbers. Still there are some lacunas in analysis techniques and algorithms because of noise data and redundancy in reads. Filling such lacuna and working on more and more efficient and computational cost effective algorithm is current necessity.



*** High Throughput Sequencing**

Figure 1: Metagenomic data analysis summery

References

1. Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Rachna, J. R., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. and Banfield, J. F. (2004): Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment. *Nature,* 428: 37-43.
2. Hugenholtz, P. (2002): Exploring Prokaryotic Diversity in the Genomic Era. *Genome Biology,* 3(2): 1-8.
3. Mardis, E. R. (2008): Next-Generation DNA Sequencing Methods. *Annu. Rev. Genomics Hum. Genet.,* 9: 387-402.
4. Steele, H. L, Streit W. R. (2005): Metagenomics: Advances in Ecology and Biotechnology. *FEMS Microbiology Letters,* 247:105–111.
5. Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R. and Wittwer, P. (2005): Metagenomic Gene Discovery: Past, Present and Future. *Trends in Biotechnology,* 23(6): 321-329.
6. Warren, R. L., Sutton, G. G., Jones, S. J. M. and Holt, R. A. (2007): Assembling Millions of Short DNA Sequences Using SSAKE, *Bioinformatics,* 23(4): 500-501.
7. Abe, T., Sugawara, H., Kinouchi, M., Kanaya, S., and Ikemura, T. (2005): Novel Phylogenetic Studies of Genomic Sequence Fragments Derived from Uncultured Microbe Mixtures in Environmental and Clinical Samples. *DNA Research*, 12: 281-190.
8. Yang, B., Peng, Y., Leung, H. C., Yiu, S., Chen, J., Chin, F. Y. (2010): Unsupervised Binning of Environmental Genomic Fragments Based on an Error Robust Selection of l-mers. *BMC Bioinformatics*, 11(S2): 1-11.

9. Mentel, M., Martin, W. (2010): Anaerobic Animals from an Ancient, Anoxic Ecological Niche. *BMC Biology*, 8:32-37.
10. Kennedy, j., Marchesi, J. R. and Dobson, A. D. (2008): Marine Metagenomics: Strategies for the Discovery of Novel Enzymes with Biotechnological Applications from Marine Environments. *Microbial Cell Factories*, 7:27-34.
11. Spencer, D. H., Kas, A., Smith, E. E., Raymond, C. K., Sims, E. H., Hastings, M., Burns, J. L., Kaul, R. and Olson, M. V. (2003): Whole-Genome Sequence Variation among Multiple Isolates of *Pseudomonas aeruginosa*. *Journal of Bacteriology*, 185(4): 1316-1325.
12. Kurikka, L. K., Lyra, A., Malinen, E., Aarnikunnas, J., Tuimala, J., Paulin, L., Mäkivuokko, H., Kajander, K. and Palva, A. (2009): Microbial Community Analysis Reveals High Level Phylogenetic Alterations in the Overall Gastrointestinal Microbiota of Diarrhoea-Predominant Irritable Bowel Syndrome Sufferers. *OMC Gastroenterology,* 9:95-105.
13. Warren, A. S., Archuleta, A., Feng, W., Setubal, J. C. (2010): Missing Genes in the Annotation of Prokaryotic Genomes. *BMC Bioinformatics*, 11:131-142.
14. White, J. R., Navlakha, S., Nagarajan, N., Ghodsi, M. R., Kingsford, C., Pop, M. (2010): Alignment and Clustering of Phylogenetic Markers - Implications for Microbial Diversity Studies. *BMC Bioinformatics,* 11:152-161.
15. Sun, and Caetano-Anolles (2010): The Ancient History of the Structure of Ribonuclease P and the Early Origins of Archaea. *BMC Bioinformatics*, 11:153.
16. Niu, *et al*. (2010): Artificial and Natural Duplicates in Pyrosequencing Reads of Metagenomic Data. *BMC Bioinformatics*, 11:187.
17. Gupta and Mathews (2010): Signature Proteins for the Major Clades of Cyanobacteria. *BMC Evolutionary Biology,* 10:24-43.
18. Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., Saar, M. O., Alexande,r S., Alexander, E. C. Jr. and Rohwer, F. (2006): Using Pyrosequencing to Shed Light on Deep Mine Microbial Ecology. *BMC Genomics,*7:57-69.
19. Hoff, K. J. (2009): The Effect of Sequencing Errors on Metagenomic Gene Prediction. *BMC Genomics,* 10:520-528.
20. Pasic, L., Rodriguez-Mueller, B., Martin-Cuadrado, A., Mira, A., Rohwer, F. and Rodriguez-Valera, F. (2009): Metagenomic Islands of Hyperhalophiles: The Case of *Salinibacter ruber*, *BMC Genomics*, 10:570-580.
21. Isaam, S. and Saman, K H. (2009): The Oligonucleotide Frequency Derived Error Gradient and its Application to The Binning of Metagenome Fragments, *BMC Genomics,* 10(3):510-522.
22. Lozupone, C., Hamady, M. and Knight, R. (2006): UniFrac – An Online Tool for Comparing Microbial Community Diversity in a Phylogenetic Context. *BMC Bioinformatics*, 7:371-384.
23. Liang, C., Schmid, A., Lopez-Sanchez, M. J., Moya, A. , Gross, R., Bernhardt, J. and Dandekar, T. J. (2009): Efficient Mapping of Prokaryotic Ests and Variable Length Sequence Reads on Related Template Genomes, *BMC Bioinformatics,* 10:391-413.
24. Gerlach, W., Unemann, S., Tille, F., Goesmann, A. and Stoye, J. (2009): WebCARMA: A Web Application for the Functional and Taxonomic Classification of Unassembled Metagenomic Reads. *BMC Bioinformatics*, 10:430-439.
25. Duncan, *et al*. (2010): DraGnET: Software for Storing, Managing and Analyzing Annotated Draft Genome Sequence Data. *BMC Bioinformatics*, 11:100-112.
26. Hyatt, *et al*. (2010): Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinformatics,* 11:119-129.
27. Mitra, S., Schubach, M. and Huson, D. H. (2010): Short Clones or Long Clones? A Simulation Study on the Use of Paired Reads in Metagenomics. *BMC Bioinformatics*, 11(1):512-522.
28. Aziz, R. K. (2010): Subsystems-Based Servers for Rapid Annotation of Genomes and Metagenomes. *BMC Bioinformatics*, 11(4): 2-3
29. Maumus, F., Allen, A. E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M. and Bowler, C. (2009): Potential Impact of Stress Activated Retrotransposons on Genome Evolution in a Marine Diatom. *BMC Genomics*, 10:624-62.
30. Grossetete *et al*. (2010): Fungi path: A Tool to Assess Fungal Metabolic Pathways Predicted by Orthology. *BMC Genomics*, 11: 81-95.
31. Tamames *et al*. (2010): Environmental Distribution of Prokaryotic Taxa. *BMC Microbiology,* 10: 85-98.