A SHORT NOTE-METAGENOMICS

Shrikant Sharma[1], Shashank Rana[1], Raghvendar Singh[2]

[1] School of Life Science, Singhania University, Jhunjhunu (RJ)
[2] Department of Biochemistry, National Research Centre on Camel, Bikaner (RJ)
***Corresponding Author:** shribioinfo@gmail.com

This article is available online at www.ssjournals.com

ABSTRACT
Metagenomics may be defined as a study of uncultured microorganisms. It includes quicker, cheaper sequencing skill and by using metagenomic approach, we can able to sequence uncultured microbial samples from their environment directly are expanding and transforming our view of the microbial world. Purify meaningful information from the millions of new genomic sequences presents a serious challenge to bioinformaticians. In cultured microbes, the genomic data come from a single clone, making sequence assembly and annotation tractable. In metagenomics, the data come from heterogeneous microbial communities, sometimes containing more than 10,000 species, with the sequence data being noisy and partial. From sampling, to assembly, to gene calling and function prediction, bioinformatics faces new demands in interpreting voluminous, noisy, and often partial sequence data. Although metagenomics is a relative newcomer to science, the past few years have seen an explosion in computational methods applied to metagenomic based research. This article gives an idea about some bioinformatic techniques for metagenomics

Keywords: Metagenomics, Microorganism, Computational Methods, bioinformatics

1. Introduction:

Metagenomics is the study of metagenomes, genetic material recovered directly from environmental samples. The broad field may also be referred to as environmental genomics, ecogenomics or community genomics. While traditional microbiology and microbial genome sequencing rely upon cultivated clonal cultures, early environmental gene sequencing cloned specific genes (often the 16S rRNA gene) to produce a profile of diversity in a natural sample. Such work revealed that the vast majority of microbial biodiversity had been missed by cultivation-based methods[11]. Recent studies use "shotgun" Sanger sequencing or massively parallel pyrosequencing to get largely unbiased samples of all genes from all the members of the sampled communities[7]. Because of its power to reveal the previously hidden diversity of microscopic life, metagenomics offers a powerful lens for viewing the microbial world that has the potential to revolutionize understanding of the entire living world[17, 18].

In the collective genomes (the metagenome) of the microorganisms inhabiting the Earths diverse environments is written the history of life on this planet. New molecular tools developed and used for the past 15 years by microbial ecologists are facilitating the extraction, cloning, screening, and sequencing of these genomes. This approach allows microbial ecologists to access and study the full range of microbial diversity, regardless of our ability to culture organisms and provides an unprecedented access to the breadth of natural products that these genomes encode. However, there is no way that the mere collection of sequences, no matter how expansive, can provide full coverage of the complex world of microbial metagenomes within the foreseeable future. Furthermore, although it is possible to fish out highly informative and useful genes from the sea of gene diversity in the environment, this can be a highly tedious and inefficient procedure. Metagenomics include technological advances in sequencing and cloning methodologies, as well as improvements in annotation and comparative sequence analysis. More significant, however, will be ways to focus in on various subsets of the metagenome that may be of particular relevance, either by limiting the target community under study or improving the focus or speed of screening procedures[8].

The term "metagenomics" was first used by Jo Handelsman, Jon Clardy, Robert M. Goodman, and others, and first appeared in publication in 1998. The term metagenome referenced the idea that a collection of genes sequenced from the environment could be

analyzed in a way analogous to the study of a single genome. The exploding interest in environmental genetics, along with the buzzword-like nature of the term, has resulted in the broader use of metagenomics to describe any sequencing of genetic material from environmental (i.e. uncultured) samples, even work that focuses on one organism or gene. Recently, Kevin Chen and Lior Pachter (researchers at the University of California, Berkeley) defined metagenomics as "the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species[4].

Conventional sequencing begins with a culture of identical cells as a source of DNA. However, early metagenomic studies revealed that there are probably large groups of microorganisms in many environments that cannot be cultured and thus cannot be sequenced. These early studies focused on 16S ribosomal RNA sequences which are relatively short, often conserved within a species, and generally different between species. Many 16S rRNA sequences have been found which do not belong to any known cultured species, indicating that there are numerous non-isolated organisms out there.

Early molecular work in the field was conducted by Norman R. Pace and colleagues, who used PCR to explore the diversity of ribosomal RNA sequences[16]. The insights gained from these breakthrough studies led Pace to propose the idea of cloning DNA directly from environmental samples as early as 1985. This led to the first report of isolating and cloning bulk DNA from an environmental sample, published by Pace and colleagues in 1991 while Pace was in the Department of Biology at Indiana University. Considerable efforts ensured that these were not PCR false positives and supported the existence of a complex community of unexplored species. Although this methodology was limited to exploring highly conserved, non-protein coding genes, it did support early microbial morphology-based observations that diversity was far more complex than was known by culturing methods.

Soon after that, Healy reported the metagenomic isolation of functional genes from "zoolibraries" constructed from a complex culture of environmental organisms grown in the laboratory on dried grasses in 1995. After leaving the Pace laboratory, Ed DeLong continued in the field and has published work that has largely laid the groundwork for environmental phylogenies based on signature 16S sequences, beginning with his group's construction of libraries from marine samples[26].

Recovery of DNA sequences longer than a few thousand base pairs from environmental samples was very difficult until recent advances in molecular biological techniques, particularly related to constructing libraries in bacterial artificial chromosomes (BACs), provided better vectors for molecular cloning[2].

Advances in bioinformatics, refinements of DNA amplification, and proliferation of computational power have greatly aided the analysis of DNA sequences recovered from environmental samples. These advances have enabled the adaptation of shotgun sequencing to metagenomic samples. The approach, used to sequence many cultured microorganisms as well as the human genome, randomly shears DNA, sequences many short sequences, and reconstructs them into a consensus sequence. Shotgun sequencing and screens of clone libraries reveal genes present in environmental samples. This provides information both on which organisms are present and what metabolic processes are possible in the community. This can be helpful in understanding the ecology of a community, particularly if multiple samples are compared to each other[1]. Shotgun metagenomics also is capable of sequencing nearly complete microbial genomes directly from the environment[29]. Because the collection of DNA from an environment is largely uncontrolled, the most abundant organisms in an environmental sample are most highly represented in the resulting sequence data. To achieve the high coverage needed to fully resolve the genomes of underrepresented community members, large samples, often prohibitively so, are needed. On the other hand, the random nature of shotgun sequencing ensures that many of these organisms will be represented by at least some small sequence segments. Due to the limitations of microbial isolation methods, the vast majority of these organisms would go unnoticed using traditional culturing techniques.

In 2002, Mya Breitbart, Forest Rohwer, and colleagues used environmental shotgun sequencing to show that 200 liters of seawater

contains over 5000 different viruses. Subsequent studies showed that there are >1000 viral species in human stool and possibly a million different viruses per kilogram of marine sediment, including many bacteriophages. Essentially all of the viruses in these studies were new species. In 2004, Gene Tyson, Jill Banfield, and colleagues at the University of California, Berkeley and the Joint Genome Institute sequenced DNA extracted from an acid mine drainage system[29]. This effort resulted in the complete, or nearly complete, genomes for a handful of bacteria and archaea that had previously resisted attempts to culture them. It was now possible to study entire genomes without the biases associated with laboratory cultures[12].

Much of the interest in metagenomics comes from the discovery that the vast majority of microorganisms had previously gone unnoticed. Traditional microbiological methods relied upon laboratory cultures of organisms. Surveys of ribosomal RNA (rRNA) genes taken directly from the environment revealed that cultivation based methods find less than 1% of the bacteria and archaea species in a sample[11].

Beginning in 2003, Craig Venter, leader of the privately-funded parallel of the Human Genome Project, has led the Global Ocean Sampling Expedition (GOS), circumnavigating the globe and collecting metagenomic samples throughout the journey. All of these samples are sequenced using shotgun sequencing, in hopes that new genomes (and therefore new organisms) would be identified. The pilot project, conducted in the Sargasso Sea, found DNA from nearly 2000 different species, including 148 types of bacteria never before seen[30] Venter has circumnavigated the globe and thoroughly explored the West Coast of the United States, and completed a two-year expedition to explore the Baltic, Mediterranean and Black Seas. Analysis of the metagenomic data collected during this journey revealed two groups of organisms, one composed of taxa adapted to environmental conditions of 'feast or famine', and a second composed of relatively fewer but more abundantly and widely distributed taxa primarily composed of plankton[32].

Using comparative gene studies and expression experiments with microarrays or proteomics researchers can piece together a metabolic network that goes beyond species boundaries. Such studies require detailed knowledge about which versions of which proteins are coded by which species and even by which strains of which species. Therefore, community genomic information is another fundamental tool (with metabolomics and proteomics) in the quest to determine how metabolites are transferred and transformed by a community[14].

Metagenomics can improve strategies for monitoring the impact of pollutants on ecosystems and for cleaning up contaminated environments. Increased understanding of how microbial communities cope with pollutants is helping assess the potential of contaminated sites to recover from pollution and increase the chances of bioaugmentation or biostimulation trials to succeed[8].

Recent progress in mining the rich genetic resource of non-culturable microbes has led to the discovery of new genes, enzymes, and natural products. The impact of metagenomics is witnessed in the development of commodity and fine chemicals, agrochemicals and pharmaceuticals where the benefit of enzyme-catalyzed chiral synthesis is increasingly recognized[32].

A major problem with metagenomes is binning. Binning is the process of identifying from what organism a particular sequence has originated. Traditionally, BLAST is a method used to rapidly search for similar sequences in existing public databases. More advanced methods have been employed to bin sequences. Big successes have been achieved for a family of methods using intrinsic features of the sequence, such as oligonucleotide frequencies. These methods include TETRA[28], Phylopythia[19], TACOA[6], PCAHIER[33], DiScRIBinATE[27] and SPHINX[21]. In 2007, Daniel Huson and Stephan Schuster developed and published the first stand-alone metagenome analysis tool, MEGAN, which can be used to perform a first analysis of a metagenomic shotgun dataset. This tool was originally developed to analyze the metagenome of a mammoth sample[25]. However in a recent study by Monzoorul et al. 2009, it was shown that adopting the LCA approach (of MEGAN) solely based on bit-score of the alignment leads to a number of false positive assignments especially in the context of metagenomic sequences originating from new organisms. This study proposed a new approach called Sort-ITEMS which used

several alignment parameters to increase the accuracy of assignments.

In 2007, Folker Meyer and Robert Edwards and a team at Argonne National Laboratory and the University of Chicago released the Metagenomics RAST server (MG-RAST) a community resource for metagenome data set analysis[20]. As of October 2011 3.7 Terabases (10^12 bases) of DNA have been analyzed by MG-RAST, more than 4300 public data sets are freely available for comparison within MG-RAST. Over 7000 users now have submitted a total of 38,000 metagenomes to MG-RAST. The server also acts as the de-fact repository for metagenomics data.

Metagenomics can improve strategies for monitoring the impact of pollutants on ecosystems and for cleaning up contaminated environments. Increased understanding of how microbial communities cope with pollutants is helping assess the potential of contaminated sites to recover from pollution and increase the chances of bioaugmentation or biostimulation trials to succeed[8].

Recent progress in mining the rich genetic resource of non-culturable microbes has led to the discovery of new genes, enzymes, and natural products. The impact of metagenomics is witnessed in the development of commodity and fine chemicals, agrochemicals and pharmaceuticals where the benefit of enzyme-catalyzed chiral synthesis is increasingly recognized[31].

2. Discussion:

Metagenomic sequencing is being used to characterize the microbial communities from 15-18 body sites from at least 250 individuals. This is part of the Human Microbiome initiative with primary goals to determine if there is a core human microbiome, to understand the changes in the human microbiome that can be correlated with human health, and to develop new technological and bioinformatics tools to support these goals [22].

It is well known that the vast majority of microbes have not been cultivated. Functional metagenomics strategies are being used to explore the interactions between plants and microbes through cultivation-independent study of the microbial communities[5].

Metagenomic sequencing is being used to characterize the microbial communities from 15-18 body sites from at least 250 individuals. This is part of the Human Microbiome initiative with primary goals to determine if there is a core human microbiome, to understand the changes in the human microbiome that can be correlated with human health, and to develop new technological and bioinformatics tools to support these goals [22].

It is well known that the vast majority of microbes have not been cultivated. Functional metagenomics strategies are being used to explore the interactions between plants and microbes through cultivation-independent study of the microbial communities[5].

Summary:

Finally, metagenomic sequencing is particularly useful in the study of viral communities. As viruses lack a shared universal phylogenetic marker (as are 16S RNA for bacteria and archaea, and 18S RNA for eukarya), the only way to access the genetic diversity of the viral community from an environmental sample is through metagenomics. Viral metagenomes (also called viromes) should thus provide more and more information about viral diversity and evolution[15].

References

1. Allen, EE, Banfield, JF (2005). Community genomics in microbial ecology and evolution. Nature Reviews Microbiology 3 (6): 489–498.
2. Beja, O., Suzuki, MT, Koonin, EV, Aravind, L, Hadd, A, Nguyen, LP; Villacorta, R; Amjadi, M et al. (2000). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environmental Microbiology 2 (5): 516–29.
3. Breitbart, M; Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002). Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy USA 99 (22): 14250–14255.
4. Chen, K and Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comp Biol 1 (2): 24.
5. Charles T (2010). The Potential for Investigation of Plant-microbe Interactions Using Metagenomics Methods. Metagenomics: Theory, Methods and Applications. Caister Academic Press. ISBN 978-1-904455-54-7.
6. Diaz NN, et al. TACOA: taxonomic classification of environmental genomic

fragments using a kernelized nearest neighbor approach. BMC Bioinformatics, 10:56.
7. Eisen, J. A. (2007). Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes.. PLoS Biology 5 (3): e82.
8. George I et al. (2010). Application of Metagenomics to Bioremediation. Metagenomics: Theory, Methods and Applications. Caister Academic Press. ISBN 978-1-904455-54-7.
9. Handelsman, J., Rondon, M.R., Brady, S. F., Clardy, J., Goodman, R., M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 5: 245–249.
10. Healy, FG; RM Ray, HC Aldrich, AC Wilkie, LO Ingram, KT Shanmugam (1995). Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. Appl. Microbiol Biotechnol. 43 (4): 667–74.
11. Hugenholz, P; Goebel BM, Pace NR (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol 180 (18): 4765–4774.
12. Hugenholz, P (2002). Exploring prokaryotic diversity in the genomic era. Genome Biology 3: 1–8.
13. H. Huson, original design by D. H. Huson and S.C. Schuster, with contributions from S. Mitra, D.C. Richter, P. Rupek, H.-J. Ruscheweyh and N. Weber (2007). MEGAN 4 - MEtaGenome Analyzer http://ab.inf.uni-tuebingen.de/software/megan/
14. Klitgord, N.; Segrè, D. (2011). Ecosystems biology of microbial metabolism. Current Opinion in Biotechnology 22 (4): 541–546.
15. Kristensen, DM; Mushegian AR, Dolja VV, Koonin EV (2009). New dimensions of the virus world discovered through metagenomics. Trends in Microbiology 18 (1): 11–19.
16. Lane, DJ; Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proceedings of the National Academy of Sciences 82 (20): 6955–9.
17. Marco, D, ed (2010). Metagenomics: Theory, Methods and Applications. Caister Academic Press. ISBN 978-1-904455-54-7.
18. Marco, D ed (2011). Metagenomics: Current Innovations and Future Trends. Caister Academic Press. ISBN 978-1-904455-87-5.
19. McHardy,A.C. et al. (2007). Accurate phylogenetic classification of variable-length dna fragments. Natural Methods, 4, 63–72.
20. Meyer, F, Paarmann D, Souza M. D, Olson R.,. Glass E. M, Kubal M., Paczian T, Stevens R., Wilke A., Wilkening, J., Edwards, R. A. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 19(9): 386.
21. Monzoorul Haque Mohammed, Tarini Shankar Ghosh, Nitin Kumar Singh and Sharmila S. Mande (2011) SPHINX—an algorithm for taxonomic binning of metagenomic sequences**,** bioinformatics, 27 (1),22–30.
22. Nelson KE and White BA (2010). Metagenomics and Its Applications to the Study of the Human Microbiome. Metagenomics: Theory, Methods and Applications. Caister Academic Press. ISBN 978-1-904455-54-7.
23. Pace, NR; DA Stahl, DJ Lane, GJ Olsen (1985). Analyzing natural microbial populations by rRNA sequences. ASM News 51: 4–12.
24. Pace, NR; Delong, EF; Pace, NR (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. Journal of Bacteriology 173 (14): 4371–4378.
25. Poinar, H N, Schwarz C, Ji Qi, Shapiro B, MacPhee R D. E., Buigues B, Tikhonov A, Daniel H. Huson, Lynn P. Tomsho, Alexander Auch, Markus Rampp, Webb Miller and Stephan C. Schuster,(2006). Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA, Science, 311(5759), 392-394
26. Stein, JL; TL Marsh, KY Wu, H Shizuya, EF DeLong (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. Journal of Bacteriology 178 (3): 591–599.

27. Tarini S Ghosh, Monzoorul Haque M and Sharmila S Mande (2010)DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. BMC Bioinformatics, 11(7):S14. http://www.biomedcentral.com/1471-2105/11/S7/S14
28. Teeling G., Jost Waldmann, Thierry Lombardot, Margarete Bauer and Frank O Glöckner (2004), TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics, 5:163-170. http://www.biomedcentral.com/1471-2105/5/163
29. Tyson, GW; Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004). Insights into community structure and metabolism by reconstruction of microbial genomes from the environment. Nature 428 (6978): 37–43.
30. Venter, JC; Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y, Smith HO (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science 304 (5667): 66–74.
31. Wong D (2010). Applications of Metagenomics for Industrial Bioproducts. Metagenomics: Theory, Methods and Applications. Caister Academic Press. ISBN 978-1-904455-54-7.
32. Yooseph, Shibu; Kenneth H. Nealson, Douglas B. Rusch, John P. McCrow, Christopher L. Dupont, Maria Kim, Justin Johnson, Robert Montgomery, Steve Ferriera, Karen Beeson, Shannon J. Williamson, Andrey Tovchigrechko, Andrew E. Allen, Lisa A. Zeigler, Granger Sutton, Eric Eisenstadt, Yu-Hui Rogers, Robert Friedman, Marvin Frazier, J. Craig Venter (2010-11-04). Genomic and functional adaptation in surface ocean planktonic prokaryotes. Nature 468 (7320): 60-66.
33. Zheng H, Wu H. (2010).Short prokaryotic DNA fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. Journal of Bioinformatics and Computational Biology, 8(6):995-1011.