Research Article

Open Access

CrossMark

Determination of uncultured microorganisms based on metagenomics signature or biomarkers at fresh water reservoirs

Abstract

Metagenomics is the branch of science which does the direct analysis of genetic material of environmental samples. Maximum microorganisms present in the ecosystem are unculturable. Metagenomics with the expansion in the field of Next Generation Sequencing (NGS) has generated large amount of raw sequence reads of uncultured microbes from different ecosystem. This combination opens door for new metabolites, drug targets, new species and unknown world of new genes could be utilized for the therapeutics. Biomarkers could be utilized as an indicator for the identification of any biological processes, disease genes, metabolites, enzymes, and antibiotics while biosensor is an analytical devise for the detection of analyte with the help of biological component (biomarkers) and physicochemical detector. Current study tried to find probable biomarkers which could be utilized for biosensor development especially for uncultured microorganisms from fresh water ecosystems. We collected all fresh water raw reads from publically available different repositories. After preliminary screening on huge data sets, we performed quality analysis on raw read sequences, generated 10 different conserve domains by multiple sequence alignment. These CDs could be promising biomarkers for the identification of uncultured microbes from fresh water ecosystems which may be pathogenic for humans. DNA hybridization based Biosensor could be synthesized by using by using this biomarker sequences for detecting the presence of uncultured microorganisms (pathogenic/non-pathogenic) in fresh water ecosystem.

Keywords: metagenomics, next generation sequencing, fresh water ecosystem, multiple sequence alignment, conserve domains, biomarkers, biosensor

Volume 2 Issue 1 - 2018

Pradnya S Panchal, Harshada V Bhairavkar, Pritee Chunarkar Patil
Rajiv Gandhi Institute of IT and Biotechnology, Bharati Vidyapeeth Deemed University, India

Correspondence: Pritee Chunarkar Patil, Rajiv Gandhi Institute of IT and Biotechnology, Bharati Vidyapeeth Deemed University, Pune, Maharashtra, India, Tel 9730038142, Fax +91-20-24365713, Email preeti.chunarkar@bharatividyapeeth.edu

Received: November 14, 2017 | **Published:** January 10, 2018

Abbreviations: NGS, next generation sequencing; MSA, multiple sequence alignment; CDs, conserve domains; EBI, european bioinformatics institute; NCBI, national center for biotechnology information

Introduction

Most of the microorganisms present in the ecology are unculturable which are commonly known as metagenomes. Metagenomics can simply define as a direct genetic analysis of this metagenome from environmental samples. It's a novel technology which can be used for exploitation of unculturable microorganisms. Metagenomics is a robust tool for discovering new hypotheses of microbial function, e.g, discoveries of proteorhodopsin based photo-heterotrophy or ammonia oxidizing Archae.[1,2] As these microorganisms are unculturable, sequencing is the major aspect of metagenomics analysis. Advances in the field of Next Generation Sequencing revolutionized the field of Biotechnology and decrease in its cost has led to the development of Metagenome Sequencing. The low cost of sequencing enables scientists to sequence large number of genetic material extracted from environmental samples. The shotgun sequences generated after sequencing can be used to determine presence of unculturable microorganisms in the environment and also their function and interaction with the environment.[3,4] This can be done with the use of conserve domains present in the sequences. Conserve domains are the sequences which remain same in related organisms i.e., they are evolutionary conserved.[5]

Now a days we don't need to rely on whole genomic sequences, only 16S rRNA sequences or 18S rRNA sequences are enough to identify bacterial or fungal species.[6] As per the data available, genomic sequences or 16S rRNA sequences or 18S rRNA sequences could be utilized for the Multiple Sequence Alignment (MSA) which can detect homology between these sequences. Multiple Sequence Alignment of nucleotide or protein sequence is most important technique in the field of computational biology and bioinformatics.[7] The conserve domains obtained after the multiple sequence alignment can be used as a marker sequences to identify the presence of microorganisms in water sample, but the CDs should be very unique to identify them. These CDs may be any specific gene as in case of species specific identification very well discussed in the paper of Yan.[8] There are biomarkers which could be utilized for the identification of more than one speciesas in the case of *Salmonella* and *E. Coli* serotyping where genes that encode the O antigen flippase *(wzx)* and the O antigen polymerase *(wzy)* are used as biomarker genes and could be utilized as biomarker for *Cronobacter* spp. as well.[9,10] This project targeted the fresh water metagenomics sequences already submitted to various metagenomics repositories. All these were genomic sequences of all microorganisms present in those environmental samples. We have generated conserve domains by performing multiple sequence alignment of the sequences. In future, by considering these sequences as a biomarker sequences for uncultured microorganisms, we can synthesize a biosensor. Biosensor is an analytical device which involves biological sensing element with wide range of application in drug discovery, food safety and environment monitoring.[11,12] It will detect the presence of uncultured microorganisms in fresh water ecosystem which may be pathogenic for humans.

Experiment design: From genome extraction to sequence storage repositories

The first step of any of such experiment is sample collection. Samples are collected in sterile condition directly from environment like water (fresh water, sea water, and sewage water etc.), aerosol, soil, serum, bio-film etc. Then the most important step in this study is extraction of DNA from the sterile sample. The protocol to extract DNA form sample will depend upon the microbial community targeted in the research. Adequate amount of DNA must be generated for further library preparation. Size fractionation also plays important role in metagenome analysis. Before Next Generation Sequencing (NGS) technology, Sanger sequencing had been using for sequencing of the genome but it was labor intensive hence now a day's NGS technology is been widely used in metagenomics analysis.[2] After ABI Solid, Roche 454 and Illumina are the two most extensively used sequencing platforms. Selection of particular NGS platform for sequencing the extracted genome has to be made on the basis of varying features of platforms like read length, degree of automation, throughput per run, data quality, and ease in analysis and cost per run etc. The main drawback of using NGS technology is the terabyte size data files generated after each run of instrument. Each run of instrument increases the computer resources requirements of sequencing laboratories.

The reads generated after sequencing are shorter compared to the reads obtain by Sanger Sequencing. Also they have origin from different organisms. Hence Assembly of the reads and data analysis of metagenome is extremely challenging process.[2,3] Assembling the generated reads into contigs can be classified into two ways – Reference based assembly and De novo assembly. The choice of route to follow depends on dataset that needs to analyze. Reference based assembly make use of reference genome as a map to create contigs using tools like MIRA while De novo assembly refers to the generation of assembled contigs using no prior reference to known genome using de-Bruijn graphs and tools such as Abyss.[2] After assembling the contigs, functional annotation, binning of sequences, variant analysis, gene/ORF prediction, community taxonomic profile, and metabolic reconstruction are the most critical steps which decide the outcome of any investigation. This entire procedure is well explained in detail in Panchal and Chunarkar-Patil review.[13]

Material and methods

Input data

European Bioinformatics Institute (EBI) Metagenomics and National Center for Biotechnology Information (NCBI) Metagenomics are two freely available resources for the study of Metagenomics and Metatranscriptomics. They provide Metagenomics sequence data derived from a range of platforms which includes Roche 454, Ion Torrent and Illumina Sequence.[14] EBI Metagenomics divides their sequence data in different biomes from which they have isolated. It includes Soil, Marine, Forest, Grassland, Freshwater, Human digestive System etc. The sequences present on EBI Metagenomics are well processed using different tools like SeqPrep and Trimmomatic. SeqPrep is a program which is used to merge the paired end Illumina reads that are overlapping into sinle longer read. Trimmomatic is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters. BioPython to remove sequences reads of length lesser than 100 etc.[15] Hence, in the following study, the Metagenomic sequences from 15 different projects of Freshwater Biome were used for the analysis.

Assembling raw reads

The raw reads present in FastQ format were assembled into contigs and then into scaffold using MIRA assembler. MIRA stands for Mimicking Intelligent Read Assembly. MIRA is a reference based assembly algorithm which uses reference genome to map the reads. MIRA assemble the reads gained by Sanger sequencing, Ion Torrent, Illumina Sequencing and Roche 454 Sequencing. MIRA uses High Confidence Region (HCR) of aligned read-pairs to start contigs building.[16]

Multiple sequence alignment

Multiple Sequence Alignment (MSA) is a widely used computational method for sequence analysis. There are different types of multiple sequence alignment algorithms which includes Pair wise alignment, Progressive alignment, Homology Search tool etc. ClustalW uses Progressive Alignment algorithm for multiple sequence alignment developed by Thompson and et al. Progressive alignment is a series of pair wise alignments to align sequences by following Neighbor Joining algorithm. ClustalW uses position specific scoring scheme and weighting scheme with 'W' representing 'Weights'. The alignment is constructed by aligning closely related sequences produce by NJ method[17,18] (Figure 1) (Figure 2).

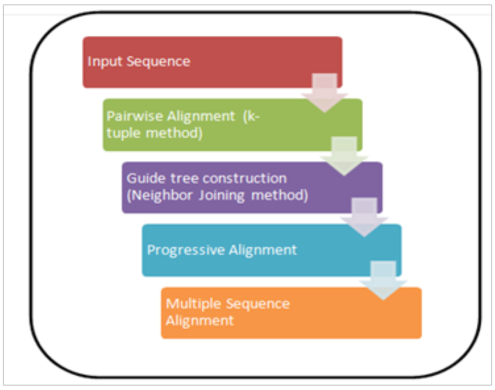
Input Sequence
Pairwise Alignment (k-tuple method)
Guide tree construction (Neighbor Joining method)
Progressive Alignment
Multiple Sequence Alignment

Figure 1 Clustal W algorithm Workflow.

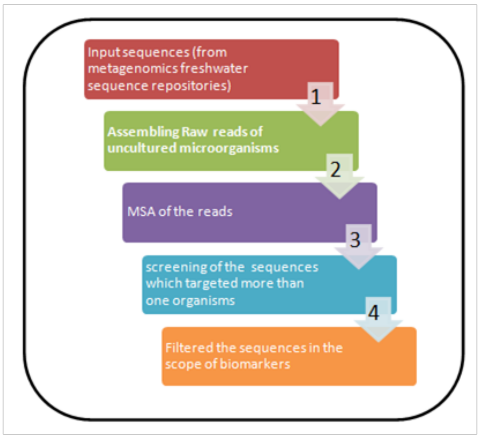
Input sequences (from metagenomics freshwater sequence repositories)
1
Assembling Raw reads of uncultured microorganisms
2
MSA of the reads
3
screening of the sequences which targeted more than one organisms
4
Filtered the sequences in the scope of biomarkers

Figure 2 Final Workflow (Step 1 to 4 has been repeated several times for the final targeted seuqnces).

Results

Taxonomical analysis

Statistical analysis of Metagenomic sequences of Freshwater Biome present from 15 different projects showed that 33% reads were unassigned. This confirms that the maximum number of sequences generated in this projects contains uncultured microorganism.[19] For the cultivable mocroorganisms, lots of work has been done on biomarkers development, even to identify the water quality aslo.[20] To identify unculturable microorganisms, lipid biomarker has been identified by Smith Carol A et.al 2000.[21] Then to date, technology has been improved wchich has ben put forth lots of data about unculturable microorganisms, but to identify them based on genetic biomarkers had not been tried. So this project tried to identify those genomic sequences from unculturable microorganisms which unable to identify them from fresh water sources.Though different projects metagenomics sequences were available, for our study, only unculturable microorganism's genomic sequences were taken into account.

Conserved domains

After performing multiple sequence alignment of the uncultured microorganism's sequences from 15 projects, we have got 10 different conserve domains which are more frequently present. These conserve domains could be used as a marker sequences to detect the presence of uncultured microbes in fresh water ecosystem. For that, step 1 to 5 has been repeated several times of the (Figure 3).

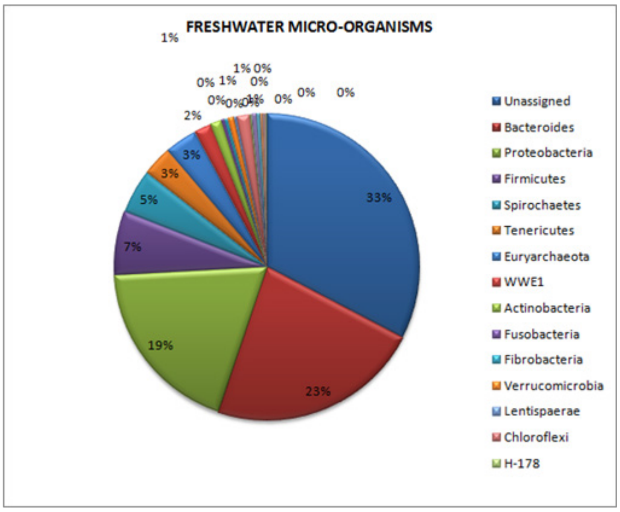

Figure 3 Percent of different types of microorganisms present in Freshwater.

Conclusion

Lots of the study has been done on the biomarkers of cultured microbial species as already discussed in the introduction part. This study has put forth the probability of finding uncultured micro-organisms biomarker sequences. There were huge sequences repositories for metagenomics sequence reads generated by high throughput sequences. Here, we only concentrate towards the fresh water bodies metagenomics sequences. The justification to select this group of sequences is to find presence of uncultured species which may be responsible for causing human diseases as fresh water bodies is utilized as drinking water in many under developed and developing countries. From all metagenomics sequences, we take only uncultured sequences for the further processing to confirm about our findings. Hence after performing multiple sequence alignment, 10 different conserve domains were observed in this metagenomic sequences which can be used to produce a biosensor which will detect the presence of unculturable microorganisms. (Table 1) also included the strength of the biomarker sequences for further utilization.

Future prospects

Biosensor is a device which has biological recognition properties for selective analysis. It is an analytical device which involves biological sensing element having applications in drug discovery, food safety, environment analysis etc. There are two different strategies used in biosensor – label based detection and label free detection. Label based detection mainly depend upon specific properties of label compounds to target detection. While Label free method – allows detecting the target molecule that are not labeled or difficult to tag.[22] The recent advance in the Biosensor has developed DNA biosensor which has opened new opportunities in research area. Hydrogel, are used as a DNA biosensor. They consider as a simple substrate for DNA immobilization. Entrapment, controlled release and DNA protection are some unique features of hydrogel.[23] A biosensor assembly includes a receptor, transducer and processor. The sensing element can be a whole cell, antibodies, enzymes or nucleic acid which integrates with transducer via immobilization.

Following types of Biosensor can be made in future to detect in presence of uncultured microorganism in fresh water.

DNA hybridization biosensor

Complementary DNA base pairing is the basis of DNA hybridization Biosensor. In DNA biosensor, 20-40 base pair long highly target selective ss-DNA segments are immobilized on electrode surface. Electric signal is produced when target DNA binds to the complementary sequence of DNA.[24]

Table 1 Conserve domains generated after multiple sequence alignment

S. no	Conserve domains
1	ACCAAC-GCCAGGGAGGGCCTGAAGAGGAGGAGGTACAGCAGCGTG-GAGTGCTACTAC-GCCGTGAGGCTGCAC Comment: Could be a strong biomarker if pattern generated as only three gaps has been introduced
2	-AAGGAGAGCGCC-TTCTACCTGAAGGAGGAG-GAGCAG--- Comment: Could be a strong biomarker if pattern generated as only three gaps has been introduced.
3	-AGCCAGAGCCTGCTGGAGTACCTGGACAAGCTGTTCGAC------------ACCGTG-------------------------ATCATC-TAC Comment: Could be a strong biomarker if domain one has only been considered.
4	--GCCATGAGGAAG-------------GGCGAGACCCTGGCCGCC TACAGGAGGAGGCTGGCCAACACCAGGGCCAGCAGGAGG--GGCAGG---- Comment: Could be a strong biomarker if domain three has only been considered.

Table Continued....

S. no	Conserve domains
5	GAGCACTACGGCGAGGCC---AGCGAC-------------CCCGGCAGGACCCACTGC----------CACCTGCTGCTGTGCCACCTGACCGAGAGC— Comment: Could be a less sensitive biomarker
6	--AGC-----ATCACCGAGCCCGAGCCCGCCTACCTGTGGCTGAAG-------------------AGGCACGTGTGGAGGTTCGTGATG--------GACATGCAC Comment: Could be a strong biomarker if domain two has only been considered
7	---------CAGCTGGAG---ACCAAGTACGTGGCCCAGCAGAAC-AGCCTGGTGGTGAACGGCTACACCGTGCCCGGC-------- Comment: Could be a strong biomarker if domain three has only been considered
8	AAGTACATCACCTACATG—AGCAAGGGCGAGTACGAGCCCGTGTTCTGCAGCAGCAGCAGCATCAGGGAC—CTGAAGCAGTGCGAC--------AAGCTGAGG— Comment: Could be a strong biomarker if domain two has only been considered
9	AAGACCATCAAGCTGAGCAAGAACACCGGCGTGCCCGTGGGCAACCCCGCCACCGGCGACACCACC-AACATCGGCATGGACAGGGTGCCC-TTCAGC Comment: Could be a best biomarker if pattern generated as only two gaps has been introduced.
10	TACAGCTGGAAGCCC-CCCAAGCTGAAGTACGGCACCGACGCCGACCTGTACCCC Comment: Could be a best biomarker if pattern generated as only one gap has been introduced.

Electrochemical DNA biosensor

Because of high sensitivity and rapid response, Electrochemical DNA Biosensor has attracted more attention. They are very useful for sequence specific bio-sensing of DNA. In electrochemical biosensor, it monitors current at fixed potential. They use both label free and labeled objects. Carbon Nanotubes also plays important role in electrochemical DNA biosensor. It helps in immobilization of DNA molecule.[25] The Biosensor to recognize uncultured microorganism can be produce by following any of the DNA biosensor hybridization technique. However, a basic research is still required to improve the sensing technology and protocol.

Acknowledgements

None.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Thomas T, Gilbert J, Meyer. Metagenomics a guide from sampling to data analysis. *Microb Inform Ex*. 2012;2(1):3.
2. Handelsman J, Rondon MR, Brady SF, et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*. 1998;5(10):245R–249R.
3. Oulas A, Pavloudi C, Polymenakou P, et al. Metagenomics: Tools and Insights for Analyzing Next–Generation Sequencing Data Derived from Biodiversity Studies. *Bioinform Biol Insights*. 2015;9:75–88.
4. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet*. 2004;38:525–552.
5. Jankun Kelly TJ, Lindeman AD, Bridges SM. Exploratory visual analysis of conserved domain on multiple sequence alignment. *BMC Bioinformatics*. 2009;10(Suppl 11):S7.
6. Woese CR, Fox GE.Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA*. 1977 Nov;74(11):5088–5090.
7. Jurate Daugelaite, Aisling O Driscoll, Roy D Sleator. An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics. Hindawi Publishing Corporation 2013:ID 615630. 2013.
8. Yan X, Gurtler J, Fratamico P, et al. Comprehensive Approaches for Molecular Biomarker Discovery for the Detection and Identification of *Cronobacter* spp. (*Enterobacter sakazakii*) and Salmonella. *Appl Environ Microbiol***.** 2011;77(5):1833–1843.
9. Fratamico PM, Yan X, Liu Y, et al. *Escherichia coli* serogroup O2 and O28ac O–antigen gene cluster sequences and detection of pathogenic *E. Coli* O2 and O28ac by PCR. *Can J Microbiol*. 2010;56(4**):**308–316.
10. Friedemann M. *Enterobacter sakazakii* in food and beverages (other than infant formula and milk powder). *Int J Food Microbiol*. 2007;116(1)**:**1–10.
11. Vigneshvar S, Sudhakumari CC, Senthilkumaran B, et al. Recent Advances in Biosensor Technology for Potential Applications–An Overview. *Front Bioeng Biotechnol*. 2016;4:11.
12. Fracchiolla NS, Artuso S, Cortelezzi A. Biosensors in clinical practice: focus on oncohematology. *Sensors (Basel)*. 2013;13:6423–6447.
13. Panchal P, Chunarkar Patil. Metagenomics: Tools and Techniques for Data Analysis. *Int J Pharma Bio Sci*. 2017;8(3):769–778.
14. Satish Kumar. *Metagenomics: Retrospect and Prospects in High Throughput Age*. Hindawi Publishing Corporation 2015: ID 121735, 2015. p. 13.
15. Mitchell A, Bucchini F, Cochrane G, et al. EBI Metagenomics in 2016–an expanding and evolving resource for analysis and archiving of Metagenomic data. *Nucleic Acid Research*. 2016;44(1):595D–603D.
16. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970;48(3):443–453.
17. Saiton N, Nei M. The neighbor–joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*. 1987;4(4):406–425.
18. Lu T, George B, Lodor M, et al. What is the Right Biomarker for Water Quality Monitoring? Pros and cons of fecal coliforms, *E. coli*, and alternative microorganisms and how they are used in watershed monitoring and water quality improvements. 2013.

19. Smith CA, Phiefer CB, Macnaughton SJ, et al. Quantitative lipid biomarker detection of unculturable microbes and chlorine exposure in water distribution system biofilms. *Water Research*. 2000;34(10):2683–2688.

20. Turner AP. Biosensors: sense and sensibility. *Chem Soc Rev*. 2013;42(3):184–3196.

21. Dias AD, Kingsley DM, Corr DT. Recent advances in bioprinting and applications for biosensing. *Biosensors (Basel)*. 2014;4(2):111–136.

22. Kavita V. DNA Biosensors–A Review. *Journal of Bioengineering & Biomedical Sciences*. 2017;7:2.

23. Arora K, Prabhakar N, Chand S, et al. Immobilization of single stranded DNA probe ontopolypyrrole–polyvinyl sulfonate for application to DNA hybridization biosensor. *Sensors Actuators B*. 2007;126:655–663.

24. Mascini M, Palchetti I, Murazza G. DNA electrochemical biosensor. *Fresenius J Anal Chem*. 2001;369(1):15–22.

25. Marrazza G, Chinella I, Mascini M. Disposable DNA electrochemical sensor for hybridization detection. *Biosens Bio–electron*. 1999;14(1):43–51.