SYNOPSIS FOR NST SCHOLARSHIP (2017-2018)

Title: Next Generation Sequencing and its Application to Phylogenetic Diversity of Metagenomics

Introduction

Next Generation Sequencing (NGS) is one of the most advanced technologies used in metagenomics studies and different computational tools have been developed for the analysis of large metagenomics dataset. NGS allows massively parallel sequencing with thousands to millions of sequences in one experiment at considerably low cost compared to Sanger method (Rastogi G (2011), Pop M (2008) and Kircher M (2010)). Advances in NGS have revolutionized the field of microbial ecology. Any NGS study follow some common steps such as sample and metadata collection, DNA extraction, library construction, sequencing and read preprocessing followed by quantitative analysis and functional binning (Metzker ML (2010) and Richter DC (2008)). Collection of environment sample is the first step in any metagenomics study. Metagenomics is a modern genomics technique which combined statistics, microbial sequence and computer technology to explore important microbes/microbial-gene associated with the complex disease/traits. The vast amount of microbial sequence data generated by NGS would be unmanageable without metagenomics tools. Without metagenomics handling, interpretation of these data would be impossible. Application of metagenomics reduces the cost of all these major projects drastically. Metagenomics is used to understand microbial gene and genome functionality. It is also used to quicken the discovery of new vaccine and design of anti-microbial agents. New metagenomics tools emerging everyday holding new promise for the improvement of agriculture and health research. Thus, traditional methods of microbial research are now enriching day by day using metagenomics. However, in Bangladesh, we are lagging behind to meet these challenges of time due to the lack of updated knowledge and expertise in metagenomics. We should discover new vaccine for any microbial diseases using the output of metagenomics tools that are adaptive to our country environment, since foreign discovery vaccine may be influenced by our environmental (Muyzer G (1993) and Cancilla MR (1992)) effect. Therefore, this has been considered on the following general objectives

Objective of the research

There are several objectives in our research; three main objectives are given below:

- Exploring better statistical approaches for improving the precision of microbial sequence assembling obtained by NGS.
- Improving statistical phylogenetic approaches for meta genome analysis to detect disease candidate microbes (virus/bacteria)
- Performance investigation of the proposed methods in a comparison of the other existing methods using both simulated and real microbial sequence analysis.

Methodology of the research

Data Sources: To investigate the performance of the proposed method in a comparison of the others existing methods using the real metagenomics data, the necessary datasets are available in the online databases, such as NCBI, SEED, MG-RAST, or COG (Wooley et al., 2010).

Statistical Methods & Software's

In this thesis I would like to explore better statistical approaches for improving the precision of microbial sequence assembling obtain by NGS. There are several NGS technique like Roche/454 genome sequencer (454 Life Sciences, Branford, CT), Illumina NGS sequencing, Ion Torrent NGS Sequencing etc. To explore the disease related microbial genes we would like to apply statistical phylogenetic tree and some statistical multivariate approaches like principal component analysis (PCA), factor analysis (FA), canonical correlation analysis (CCA) and some supervised and unsupervised clustering/classification techniques (Edwards et al., 2012).

Software: Mainly I have to use R statistical package and Primer-E package and for analytical amenities SPSS, MS Excel etc will appear assist to R. Also I would like to use the python programming for preprocessing the microbial sequence.

Time frame

The working time frame of my study is given below:

Research Activities	2017						2018					
	July	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	April	May	Jun
Review of Literature												
Development of proposed algorithm												
Simulation study												
Performance measure for real data analysis												
Thesis writing and submission												
Paper writing and submission												

Socio-economic importance

The idea of this thesis can be utilized to explore important microbial genes associated with any diseases or complex traits of any organism. So, it can also be utilized for the discovery of new vaccine in the health sector for any complex diseases of any organism as well as for the improvement in the agricultural sectors. It can contribute to the health sectors by reducing the cost and time period of discovering an appropriate vaccine for any complex diseases to improve the health care facilities for the society for any environmental conditions such as global warming because some of the microbial species and functional components involved in greenhouse gas emissions. It can also be useful for building capacities of developing high-level scientific research in metagenomics in our country. Thus the output of the proposed thesis may significantly contribute to the socio-economic development in our country.

Conclusion

In this thesis, we would like to explore better statistical approaches for microbial sequence assembling to extract the whole genome sequence more accurately. Then we will explore better statistical approaches for the analysis of whole genome sequence dataset to detect the important microbes (Virus/Bacteria) associated with any diseases or complex traits of any organism. The performance of the statistical approaches will be investigated using both

synthetic and real microbial sequence analysis. Then an attempt would be made to suggest some statistical approaches suitable for microbial sequence analysis to identify important genes associated with any diseases or complex traits.

References

1. Rastogi G, Sani R (2011) Molecular Techniques to Assess Microbial Community Structure, Function, and Dynamics in the Environment. *In: Springer* New York 29-57.
2. Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet* 24: 142-149.
3. Kircher M, Kelso J (2010) High-throughput DNA sequencing: Concepts and limitations. *Bioessays* 32: 524-536.
4. Metzker ML (2010) Sequencing technologies-the next generation. *Nat Rev Genet* 11: 31-46.
5. Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008) MetaSim-A Sequencing Simulator for Genomics and Metagenomics. *PLoS ONE* 3: e3373.
6. Muyzer G, de Waal EC, Uitterlinden AG (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction amplified genes encoding for 16S rRNA. *Appl Environ Microbiol* 59: 695-700.
7. Cancilla MR, Powell IB, Hillier AJ, Davidson BE (1992) Rapid genomic fingerprinting of Lactococcuslactis strains by arbitrarily primed polymerase chain-reaction with P-32 and fluorescent labels. *Appl Environ Microbiol* 58: 1772-1775.

Supervisor	**Student**
	(Md. Mazharul Islam)
	M.sc (Thesis)
	Roll No: 13114787
(Dr. Md. Nurul Haque Mollah)	Reg. No: 2075
Professor & Convener of Bioinformatics research group	Session: 2016-2017
Laboratory of Bioinformatics	Department of Statistics,
Department of Statistics,	University of Rajshahi,
University of Rajshahi, Rajshahi-6205	Rajshahi-6205
Phone: +8801715-319178	Phone: +8801773-371371
Email: mollah.stat.bio@ru.ac.bd	Email: shouravstat13@gmail.com