Analysis of Metagenomes of Wastewater

Hiteshree Lokhande

PG Research Scholar G H Raisoni college of Engineering Nagpur

Sonali Nimbhorkar

Assistant Professor,G H Raisoni College of Engineering Nagpur

Abstract

Metagenomic analysis process include understanding of symbiotic systems and effects of environment pollutions. However, metagenome analysis requires sequence homology searches which require large computation time and it is thus a challenge in current metagenome analysis based on the data from the latest DNA sequencers generally called a next-generation sequencer. Metagenome analysis is the understanding of the genomes of microbes which is obtained directly from microbial communities in their natural habitats such as soils, seas, and human bodies. It includes genomic sequences which is obtained directly from environmental microbial communities with the motive of linking their structures with functional roles.

The major problem of this work is to find distribution of phenol degradation of genes and tracing any possible variations in its genes across various sources. The increased antibiotic resistance among microorganisms has resulted into growing interest for investigating the wastewater treatment plants (WWTPs) as they are reported to be the major source in the dissemination of antibiotic resistance genes (ARGs) and heavy metal resistance genes (HMRGs) in the environment. Biological wastewater treatment is among the most important biotechnological applications and, as drivers of the key processes, microorganisms are central to its success. Therefore, the study of wastewater microorganisms has obvious applied significance. Metagenomes also referred to as environmental and community genomics is the genomic analysis of microorganisms by direct extraction and cloning of DNA from an assemblage of microorganisms.. Metagenome consist of genetic material of environmental samples, here the computational tools are used,using it we can study the protein sequences of different bacteria and compare their genetic information present in them and can then classify them into different bacteria according to genetic analysis.The challenges we faced is metagenomes are taken from different places,their protein structure is different from genomes,understanding their characteristics, collection of those sequences from database takes long time,so using some computational tools time taken to perform these task should be minimized.

Keywords: Metagenomics,WWTPs,ARGs, HMRGs

INTRODUCTION

Rapid development of sequencing technology and the usage of computational tools changes for metagenomic analysis which determines how the symbiotic system works and finding pollution in environment. Metagenomics is genetic information of the genes of environmental samples including wastewater. Metagenomics has shown enormous potential to drive the discovery and understanding of uncultured communities which contains various unknown microbes that cannot be cultivated in the laboratory and sequenced by traditional means. Since these sequencing technologies produces short reads from long genomes, in case of metagenomic sequencing, the task of identifying multiple species becomes more difficult. To analyze and annotate the large repository of available sequence data from different microbial communities, researchers are building different computational tools. Several algorithms have been developed to analyze targeted metagenomics .A stage in genome projects is the assembly of shotgun reads, or collecting together short DNA fragments randomly extracted from the sample, to form a set of longer contiguous stretches of DNA strings called contigs, which is the basis of many important downstream analysis e.g., gene prediction, annotation, and genomic variation. Genome assembly is one of the most computationally challenging tasks in metagenomics, and the performance of current tools are far from satisfactory. In traditional single genome projects, the major cause of misassembly is repetitive regions (DNA repeats). Despite these challenges, various computational tools, have been developed and successfully applied to the assembly of large individual genomes from massive amount of short reads. The assembly of metagenomes, however, brings about additional assembly challenges in the form of nonuniform read depth due to nonuniform distribution of species abundance, multi-strain population in the natural environment, and the potential for the coassembly of reads originating from different species.

Metagenomic analysis is the study of the genomes of uncultured microbes obtained directly from microbial communities in their natural habitats such as soils, seas, and human bodies. The analysis is useful for not only understanding symbiotic systems but also watching environment pollutions . These days, the latest DNA sequencers, generally called next-generation sequencer, became to produce huge amount of genomic data in a short time and it is expected that metagenomic researches are promoted based on such genomic data. Metagenome analysis which requires comparisons of sequence data obtained from a sequencer with sequence data of remote homologues in databases. Because the current databases do not include sequence data for most of microbes in the sample. Therefore, sensitive sequence homology search processes are required in metagenome analysis. Unfortunately, this process needs large computation time and is thus a challenge in current metagenome analysis based on the data from a next-generation sequencer. In this paper, we developed a large-scale automated computing pipeline for analyzing huge amount of metagenomic data obtained.

Metagenomics enables the study of unculturable microorganisms in different environments. Discriminating between the compositional differences of metagenomes is an important and challenging problem. Several distance functions have been proposed to estimate and find out the differences based on functional profiles,however, the strengths and limitations of such functions are still unclear. Initially, it has analyzed and found very little difference between them in the clustering of samples. This motivated us to incorporate suitable normalizations and phylogenetic information into the functions so that we could cluster samples from both real and synthetic data sets. The results proves significant improvement in sample clustering over that derived by rank-based normalization with phylogenetic information, regardless of whether the samples are from real or synthetic microbiomes. The application of metagenomes sequence information will facilitate the design of better culturing strategies to link genomic analysis with pure culture studies Furthermore, our findings indicates that considering suitable normalizations and phylogenetic information is very essential when designing distance functions for estimating the differences between metagenomes. We suggest that incorporating rank-based normalization with phylogenetic information into the distance functions helps achieve reliable clustering results.

Literature Review

[1]The Impact of Normalization and Phylogenetic Information on Estimating the Distance for Metagenomes incudes the study of Metagenomics provides the key to exploring microbial communities, and clustering metagenomes is, therefore, an important step toward revealing the relationships between microbiomes. To address this issue, we compared three well-known distance functions, namely, the Euclidean, Manhattan, and Pearson functions, and incorporated different normalizations into the functions. Our results indicate that using rank-based normalization to analyze metagenomic samples may make it easier to distinguish between microbiomes. In addition to, taking phylogenetic information into consideration further improves the clustering process.

[2] Taxonomic Classification DNA Fragment of Metagenome with a Novel

Model ,it has proposed a novel classifier to assign the unknown DNA fragments of metagenome to their taxonomic levels and taxonomic groups. In training phase, we divided each class at each taxonomic level into three cases: accept, reject and uncertain by applying the SVDD model, in order to enhance the discriminating ability of the classifier. In classification phase, we designed a score strategy to seek their optimal taxonomic level and taxonomic group for these query DNA fragments. The experimental result shown that the performance of our classifier is better than some existing classifiers with DNA fragments lengths ranging from 500 bp to 3,000 bp.

[3] A Comprehensive Study on Predicting Functional Role of Metagenomes Using Machine Learning Methods ,the current study makes important methodological contri-butions for use of ML models in the field of functional metagenomics. In this study, it has uniformly evaluated meta-genomic human microbiome data from 3 studies and used 10 folds cross-validation to evaluate the performance of ML models used for prediction of functions.It has recom-mended some of the best models in general for functional metagenomic analysis. It have shown that embedded feature selection strategies of XGBoost or Glmnet are most effective in dealing with high-dimensional metagenomic data, as they are faster, provide better performance and are scalable with large scale metagenomic data.

[4] Binning DNA Fragment of Metagenome Using a Novel Model, in this article, it has predicted unknown taxonomic organism of metagenome using SVDD classifier at genus level. The classifier can eliminate the interference from some outliers for reference genomic data, and then generate more accurate data domain description for each taxonomic class. The experimental results demonstrate that the classifier has the performance above average at genus level. Meanwhile, being an independent method, it provides valuable suggestions on taxonomic revisions for other methods. When other methods do not possess enough resolution, for example, in the case of multiple species of a genus, the SVDD method supplies additional information.

[5] An ultra-fast computing pipeline for metagenome analysis with next-generation

DNA sequencers, it includes for analyzing metagenomic data obtained from a next generation sequencer in real time, we developed a large-scaleautomated computing pipeline which enables us to utilize huge computational resources on TSUBAME2. The results of the experiment with whole TSUBMAME2 system indicate the pipeline can process genome information obtained from a single run of next-generation sequencers in a few hours.

[6] Parallel and Memory-efficient Preprocessing for Metagenome Assembly, in this work, it has developed a new bioinformatics tool called METAPREP for partitioning metagenomic reads into disjoint components, such that each component can be independently assembled by existing assemblers. Empirical results show that METAPREP exhibits good strong scaling on a single compute node with 24 cores, while also exhibiting reasonable scaling on multiple compute nodes. We also evaluated the performance of individual steps in METAPREP by comparing with corresponding stateof- the-art implementations.

[7]BCP-MG: A Web Server for Predicting Bacterial Community of Metagenome:BCP-MG is a user-friendly tool predicting the bacterial community according to the enzyme information extracted from the annotation of metagenomic sample or the metabolic network reconstructed from this data. We evaluated our method on real metagenomic data by comparison with MGRAST. We could show a close correspondence between the predicted bacterial community of BCP-MG and MG-RAST, while BCP-MG predict a conservative bacterial community which not only covers all the known metabolic functions, but also can be used to discover unknown functions of the metagenomic samples

METHODOLOGY

This study involves information about proteins ,bacterias and their protein structure.first task was to retrieve the sequence report from ncbi database.The study of metagenomes and their characteristics has been done before long time but since time grows metagenomes and is classification and its characterstics changes frequently.

Earlier, the study of metagenomes would take long time such as performing sorting retrieving and manipulating the data and the data consist of metegenomes sequences.Using entrez command which is used to retrieve data from ncbi database is used to retrieve sequence .Second task is to differentiate those bacteria which are in bracket from other bacteria.third task is to retrieve genbank file and fasta file of given bacteria from ncbi database and fourth task is to sort out different bacteria according to different unit,fifth task is to generate protein structure from PFAM database.sixth task was to generate phylogenetic tree using MEGAN tool.Earlier some python programs were developed but below ones are most efficient than earlier regarding its its complexity and consumption of time to execute those tasks.

Retrieving sequence report

To download sequence of protien and nucleotide in fasta file

Program:

```
Esearch –db protien – query phenol hydroxylase
   Efetch –format fasta
```

To take out only particular bacteria name which lie in bracket

```
infile = open('out3.txt', 'r')
outfile = open('out5.txt', 'w')
 for line in infile:
line = (line[line.find('[') + 1:])[:-2] + "\n"
outfile.write(line)
 infile.close()
outfile.close()
```

if we want to remove duplicate data then below program will work

```
  infile = open('out3.txt', 'r')
outfile = open('out6.txt', 'w')
unique = []
 for line in infile:
 if line not in unique:
unique.append(line)
outfile.write(line)
 infile.close() outfile.close()
```

In the above two programs out3.txt,out5.txt,out6.txt files are those which contain protein sequences of different bacterias of metagenomes.It consist of data retrieved from NCBI database.

It includes searching of data from ncbi database and then sorting the bacteria file using different python programs which described above.Using the sequences of different bacterias it lead to finding protein structure in PFAM database,sorting the data according to protein structure and then formed textfile containing bacteria of same type,the text file is then used in MEGAN tools to obtain megan file which is then used to generate phylogenetic tree. The tree gives the structure definition of different bacteria having same protein structure.The phylogenetic tree tells us difference about genomic bacteria and metagenomic bacteria and their protein structure.It has included searching of metagenome using MG-RAST tool ,it found out different metagenome from different places having various species and then again using PFAM database the protein structure was found out of those metagenomes,by using the sequence of hydroxylase bacteria.

Its objective was to design various program and code for solving issues of biological data retrieval sorting, manipulation and analysis and understanding various bioinformatics tools and their algorithm for analyzing genomic and metagenomic data.

As the era of genomic and metagenomic is progressing more and more data are being generated .There is need for specific program to properly analyze such data . There are thousand of bacteria with thousand of protiens in given environment,to understand the distribution of any given protien across several bacteria is a quite challenging task. This study thus focus in development of programs and taking help of other available bioinformatics tools to understand the pattern of protien distribution and evolution

The study involves learning different genomic and metagenomic bacterias,the case study involves how the protein structure of these bacteria vary according to their regions where they are located.The phylogentic tree which is formed after obtaining the protein structure and classifying them into different units tells us the information about the characteristics and properties about those metagenomic bacteria and it has came to conclusion that metagenome are totally different than genomes

The different places contain different species of metagenomes and according to their types such as terrestrial,marine,air,forest,freshwater they vary and have different structural characteristics.It is wide study to understand the metagenomes and their protein structure and the future work also involves the existences of new emerging bacteria in different places and to study the genetic information of those bacteria will be continued further.

RESULT AND DISCUSSION

To gain a better understanding of the distance functions used to estimate compositional differences between metagenomes, we started by comparing the performances of three well-known distance functions on clustering samples in a metagenomic data set. We found that applying suitable normalizations and considering the relationships among species are very important to ensure the clustering accuracy of distance functions. Therefore, we decided to consider the impact of different normalizations on the clustering results. We also incorporated the phylogenetic information about the target species into the distance functions. Tests of the functions on both real and synthetic metagenomic data sets demonstrated that the functions derived by incorporating rank normalization and phylogeny at the family level definitely yielded more reliable estimations.Comparison of Distance Functions based on Clustering Metagenomic Samples .The results of using the Euclidean, Manhattan, and Pearson functions to cluster samples of a real metagenomic data set are presented. All the fish microbiomes and six out of the seven coral microbiomes were tightly clustered by the three distance functions. In addition, the following microbiomes were grouped together by the distance functions: marine-1 and marine-3, marine-5 and marine-6 and terrestrial-animals-1, 2, and 3. We observe that freshwater-1 is always separated from the other freshwater samples irrespective of the function used in this study, which is consistent with the results reported. In spite of the similarity in sample clustering when using these distance functions, we found a small difference between the functions; terrestrial-animals-7 was clustered close to terrestrial-animals-1, 2, and 3 when the Manhattan function was used, but it was farther away when the Euclidean and Pearson functions were applied. Our results indicate that the functions yield similar performances; however, in some cases, the Manhattan function achieved slightly more reasonable clustering results. In addition, comparing the formula of the three functions led to some interesting observations. The Manhattan function is constrained to single-axis movement, and it calculates the distance under the assumption that all OTUs are independent of each other. However, it is recognized that, to understand the differences between two communities, information about the composition of the microbial community as well as it might not be robust against metagenomic samples, which are frequently populated by dominant proportions of a few species. Suitable normalizations of samples are usually required to solve this problem. Hence, we investigated whether incorporating different normalizations and considering phylogenetic information would yield a better estimation when designing distance functions for metagenomic analysis

we developed a method for estimating the possible number of species from the observed NGS read sequences of metagenomes by taking advantage of clustering and genome signatures. Our method is critical for an accurate and efficient down-stream analyses, such as metagenome assembly, and further contributes to the better understanding of microbial community in metagenomes.

CONCLUSION

To analyze metagenomic data obtained from a next generation sequencer in real time, we developed a large-scale automated computing pipeline which enables us to utilize huge computational resources. The results of the experiment with whole PFAM,MG-RASTsystem indicate the pipeline can process genome information obtained from a single run of next-generation sequencers in a few hours. It is our hope that this work will contribute to understand the classification and clustering in the field of metagenomics. This also provides theoretical foundations of our proposed methods as well as other algorithms used in computational biology research. By incorporating phylogenetic information, our methods produces more rational clustering of metagenomic samples than that derived by conventional approaches. The results also tells that the performance of phylogenetic information at the family level is better than at the phylum and species level. Nevertheless, because of the various features of metagenomic data such as community complexity, sequencing tools, and genome coverage, different methods of analysis e.g., distance functions, normalization, and phylogenetic information should be tested and compared in order to achieve optimal results. Furthermore, due to the limitations of metagenomic analysis, this as short reads, low genome coverage, bad sequencing quality, and biased annotation, discerning the differences between metagenomic samples is still a huge challenge. Improvements in sequencing technology and the completeness of new microbial genome projects in the near future should be helpful in solving the remaining analytical bottlenecks.

REFERENCES

[1] Chien-Hao Su, Tse-Yi Wang, Ming-Tsung Hsu, Francis Cheng-Hsuan Weng, Cheng-Yan Kao, Daryi Wang, and Huai-Kuang Tsai,"The Impact of Normalization and Phylogenetic Information on Estimating the Distance for Metagenomes" IEEE/ACM Transactions on computational biology and bioinformatics,March-April 2012

[2]Tao Hou1, Yun Liu1, Jian Xue1, Mingming Li2, Fu Liu11. College of Communications Engineering, Jilin University, Changchun 130022, China 2. Network center, Jilin University, Changchun 130022, China] Taxonomic Classification DNA Fragment of Metagenome with a Nove Model,2016 35 chinese control conference

[3] Jyotsna Talreja Wassan, Student Member, IEEE, Haiying Wang, Fiona Browne, Member, IEEE and Huiru Zheng*, Senior Member, IEEE A Comprehensive Study on Predicting Functional Role of Metagenomes Using Machine Learning Methods,IEEE/ACM Transaction on computational Biology and Bioinformatics,2018

[4] Hou Tao, Liu Yun, Liu Fu, Wang Ke, Xie Jian College of Communications Engineering, Jilin University, Changchun 130022 E-mail: liufu@jlu.edu.cn Binning DNA Fragment of Metagenome Using a Novel Model,27th chinese control and decision conference(2015CCDC)

[5] Shuji SuzukiTakashi IshidaYutaka AkiyamaAn ultra-fast computing pipeline for metagenome analysis with next-generation DNA sequencers,2012 SC Companion High performance computing networking storage and analysis.

[6] Vasudevan Rengasamy Paul Medvedev Kamesh Madduri
*The Pennsylvania State University University Park, PA, USA*Parallel and Memory-efficient Preprocessing for Metagenome Assembly,2017IEEE international parallel and distributed process symposium workshop

[7] BCP Deguo Xia #1, Haoran Zheng *#2, Zhiqiang Liu #3, Guisheng Li #4, Kai Zhao #5 # College of Computer Science and Technology, University of Science and Technology of China, Hefei, PR China-MG: A Web Server for Predicting Bacterial Community of Metagenome,2010 IEEE 5th conference on bioinspired computing

[8] Chien-Hao Su, Tse-Yi Wang, Ming-Tsung Hsu, Francis Cheng-Hsuan Weng, Cheng-Yan Kao, Daryi Wang, and Huai-Kuang Tsai. The Impact of Normalization and Phylogenetic Information on Estimating the Distance for Metagenomes,IEEE/ACM transactions on computational biology and bioinformatics

[9] Ho-Sik Seok, Woonyoung Hong Department of Animal Biotechnology Konkuk University Seoul, Korea Jaebum Kim Department of Animal Biotechnology and UBITA Center for Biotechnology Research Konkuk University Seoul Estimating the Number of Species in Metagenomes by Clustering Next-Generation Read Sequences,2014Internatioonal conference on Big data and smart computing

[10] Shuji Suzuki Department of Computer Science, Graduate School of Information Science and Engineering Tokyo Institute of Technology Tokyo Takashi Ishida (Adviser) Department of Computer Science, Graduate School of Information Science and Engineering Tokyo Institute of Technology Tokyo, An ultra-fast computing pipeline for metagenome analysis with next-generation DNA sequencers IEEE conferences,2012 SC companion,High Performance computing,networking storage and analysis

[11]Bo Liu and Mihai Pop Department of Computer Science Center for Bioinformatics and Computational Biology, UMIACS University of Maryland, College Park, USA {boliu, mpop}@umiacs.umd.edu Workshop: Comparative Assembly of Metagenomic Sequences, 2012 IEEE 2nd International Conference on Computational Advances in bio and medical sciences

[12] Zeehasham Rasheed Huzefa Rangwala Department of Computer Science Department of Computer Science George Mason University George Mason University Fairfax, VA 22030 USA Fairfax, VA 22030 USA Patrick Gillevet Microbiome Analysis Center Department of Environmental Science and Policy George Mason University Classification and Clustering in Metagenomics with Unified Data Management and Computational Framework,2012 IEEE International conference on bioinformatics and biomedicine workshops

[13] J. Handelsman, J. Tiedje, L. Alvarez-Cohen, M. Ashburner, I.K.O. Cann, and E.E. DeLong, The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Nat'l Academies Press, 2007. [14] S. Mitra, J.A. Gilbert, D. Field, and D.H. Huson, "Comparison of Multiple Metagenomes Using Phylogenetic Networks

Based on Ecological Indices," ISME J., vol. 4, pp. 1236-1242, Apr. 2010.

[15] E.F. DeLong, C.M. Preston, T. Mincer, V. Rich, S.J. Hallam, N.U. Frigaard, A. Martinez, M.B. Sullivan, R. Edwards, B.R. Brito, S.W. Chisholm, and D.M. Karl, "Community Genomics among Stratified Microbial Assemblages in the Ocean's Interior," Science, vol. 311, no. 5760, pp. 496-503, Jan. 2006.

[16] K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V.K. Sharma, T.P. Srivastava, T.D. Taylor, H. Noguchi, H. Mori, Y. Ogura, D.S. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi, and M. Hattori, "Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes," DNA Research, vol. 14, no. 4, pp. 169-181, Aug. 2007.

[17] W. Li, "Analysis and Comparison of Very Large Metagenomes with Fast Clustering and Functional Annotation," BMC Bioinformatics, vol. 10, article 359, 2009.

[18] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," Genome Res., vol. 17, pp. 377-386, 2007.

[19] J. C. Wooley, A. Godzik, and I. Fridberg, "A primer on metagenomics," PLoS Comput Biol., vol. 6, pp. e1000667, 2010.

[20] S. Boisvert, F. Raymond, E. Godzaridis, F. Laviolette, and J. Corbeil, "Ray Meta: scalable de novo metagenome assembly and profiling," Genome Res., vol. 13, pp. R122, 2012.