J. Bio-Sci. 24: 45-53, 2016 ISSN 1023-8654
http://www.banglajol.info/index.php/JBS/index

MULTIVARIATE STATISTICAL TECHNIQUES FOR METAGENOMIC ANALYSIS OF MICROBIAL COMMUNITY RECOVERED FROM ENVIRONMENTAL SAMPLES

Z Akond[1, 2*], M Alam[3], MS Ahmed[4] and MNH Mollah[4*]

[1]Agricultural Statistics and Information and Communication Technology Division, Bangladesh Agricultural Research Institute, Bangladesh; [2]Institute of Environmental Science, University of Rajshahi, Bangladesh; [3]Emerging Infections, Infectious Diseases Division, International Centre for Diarrheal Disease Research, Bangladesh (icddr,b); [4]Bioinformatics Lab., Department of Statistics, University of Rajshahi, Bangladesh

Abstract

High-throughput big dataset generated through next generation sequencing (NGS) of DNA samples helps identify key differences in the function and taxonomy between microbial communities as well as shed light on the diversity of microbes, cooperation and evolution in any particular ecosystem. During this study, three statistical techniques namely, Random Forest (RF), Multidimensional Scaling (MDS) and Linear Discriminant Analysis (LDA) approaches were employed for functional analysis of 212 publicly available metagenomic datasets within and between 10 environments against 27 metabolic functions. RF generates the 8 most important metabolic variables along with MDS and LDA among which Photosynthesis has the highest score (70.20); Phages, prophages has the second highest score (61.31) and Membrane Transport was found to have the eighth highest score (45.29). The MDS plot was found useful to visualize the separation of the microbes from human or animal hosts from other samples along the first dimension and the separation of the aquatic and mat communities along the second dimension. LDA analyses compared the extent of the microbial samples into three broad groups: the human and animal associated samples, the microbial mats, and the aquatic samples. RF showed that phage activity is a major difference between host-associated microbial communities and free-living. The MDS and LDA techniques suggest that mat communities were unique from both the animal associated metagenomes and the aquatic samples with differences in the vitamin and cofactor metabolism.

Key words: Environment, functional role, linear discriminant analysis, metagenomes, multiple dimensional scaling, random forest

Introduction

Most life on this planet is microbes that help to maintain the ecological balance greatly through their direct and indirect interactions with biotic and abiotic components of the environment. Metagenomics has recently started contributing to reveal the actual scenario surrounding biodiversity of this microbial life. The technique comprises extracting and sequencing the DNA and RNA (metatranscriptomics) of microbial communities collected directly from any specific samples e.g., human or plant/animal-associated, environmental, industrial, food sources and then using high performance computational and statistical analysis to associate function to each sequence (Dinsdale et al. 2013). Due to rapid advancement in IT as well as the continued and dynamic development of faster next generation sequencing technologies with various platforms, it is now a powerful tool to sequence multiple samples with millions of short DNA fragments or reads in a single run. This ultimately facilitates studying the multiple microorganisms living in an environmental community without the need of isolating and culturing individual microbial species in a laboratory. It has been reported that more than 99% out of the millions of microbial species known to exist on earth cannot be cultured in a laboratory

[*]Author for correspondence: akond25@yahoo.com

(Huson et al. 2009). However one can attempt to generate new NGS metagenomic dataset with a view to detect organisms of similar nature. Further annotation of a possible metagenome is conducted by comparing the sample DNA to the sequences that are available in various databases such as NCBI, SEED, MG-RAST, or COG (Aziz et al. 2008, Wooley et al. 2010). In most cases, the DNA sequences similar to each corresponding protein are identified; therefore a metagenome provides information on the taxonomic makeup and metabolic potential of a microbial community (Tringe et al. 2005). However, one should be aware that if new sequences are found, they might not give any hit which needs further analyses to claim novel and previously unknown organisms.

Until now, most of the focus in metagenomics has been on single environments such as coral atolls (Wogley et al. 2007, Dinsdale et al. 2008), cow intestine (Brulc et al. 2009), ocean water (Angly et al. 2006), and microbialites (Breitbart et al. 2009). Early work compared extremely different environments like soil microbes compared to water microbes (Breitbart et al. 2009). More recently, the Human microbiome project (https://hmpdacc.org) has expanded our understanding of the microbes inhabiting our own bodies, comparing samples from the same site among and between individuals (Tunrbaugh et al. 2010). These studies reflect the dynamic and expanding field of metagenomics which has been shown elsewhere (Wogley et al. 2007). Metagenomics provides a complete analysis of the microbial activity in terms of how the microbial community or metabolic potentials (of a group of organism) vary between sampling locations or at different time points (Kurokawa et al. 2008). During this study attempt has been made to explore the abilities of metagenomics technique while analyzing the metabolic profile of microbial communities with eventual visualization of large amounts of multivariate data.

Materials and Methods

A total of 212 metagenome datasets were selected from publicly available database (https://dinsdalelab.sdsu.edu/metag.stats/).These were classified into 10 different environments depending on the descriptions provided by the researcher who submitted the data. The experiments were involved a number of NGS sequencing tools (Pyrosequencing and Roche Applied Sciences and 454 Life Sciences GS20 Platforms).

Data source

The metagenomes used in this article are freely available from the SEED platform and are being made accessible from CAMERA and the NCBI Short Read Archive. The NCBI genome project IDs used in this study are: 4441143- 44, 4441148, 4441152-53, 4441579 -86, 4441589, 4441591, 4441595-97, 4441600-02, 4441605, 4441613, 4441618, 4441658-60, 4441662, 4440361-65, 4443688-89, 4443691, 4443693, 4443702-04, 4443706-09, 4443711-15, 4443718-22, 4443724-25, 4441041, 4441056-57, 4441062, 4441590, 4443679-81, 4443683-85, 4443687, 4440411, 4440413, 4440422, 4440440, 4441092, 4441093, 4440453-54, 4440461-63, 4440595, 4440610-11, 4440613-16, 4440639-40, 4440823-26, 4440939, 4440940-51, 4441050,4441599,4440324,4440329,4440416,4440419,4440425-26, 4440429-30, 4440433-35, 4440437-38, 4440963-72, 4441051,4441055,4441057,4441125-30, 4441134-36, 4441139,4441145-47, 4441149-51, 4441155-56, 4441570,4441573-78, 4441587-88,4441592,4441594,4441607,4441609-11,4441614-16,4441661,4443740,4441121,4441133,4441139,4441167,4441593, 4441603-04, 4441617, 4442642-43,4442643, 4442647-53, 4440037, 4440039-41, 4442583, 4443746-47, 4443749-50,4443750,4443762,4441679-84, 4440463-64,4440464,4440056. These processed dataset were collected from the website of Dinsdale Lab., San Diego State University (https://dinsdalelab.sdsu.edu/metag.stats/) published in 2009. A number of statistical techniques were applied to these metagenomics data to explore the relevant phenomenon (Dinsdale et al. 2013).

While various environmental measurements were collected at the time of metagenome sampling, the two data types: environmental and genomic have been analyzed simultaneously to provide direct evidence of how microbial communities differ across environmental gradients. Therefore our analyses used the percent of sequences in each metabolic or functional group as the datasets. The metabolic group is the response variables and the metagenomes were considered as the observations. The 27 functional hierarchies used in the analysis were: amino acids and derivatives; carbohydrates; cell division and cell cycle; cell wall and capsule; cofactors, vitamins, prosthetic groups and pigments; DNA metabolism; dormancy and sporulation; fatty acids, lipids, and isoprenoids; membrane transport; metabolism of aromatic compounds; miscellaneous; motility and chemotaxis; nitrogen metabolism; nucleosides and nucleotides; phages, prophages and transposable elements; phosphorus metabolism; photosynthesis; plasmids; potassium metabolism; protein metabolism; regulation and cell signaling; respiration; RNA metabolism; secondary metabolism; stress response; sulfur metabolism and virulence as classified by the concerned researchers (Aziz et al. 2008).

Statistical and Graphical methods

The data consisted of 10 different types (the environments), 27 response variables (the functional metabolic groups), and 212 observations (the metagenomes). we attempted to analyze multivariate data of the metagenomes using three different widely used statistical techniques namely random forests (RF), multidimensional scaling and linear discriminant analysis with a view to visualize the differences between and within environments and identify the key metabolic processes that might be crucial in the biological process.

Random forests

The random forest (Brieiman 2001) is a robust analytical tool. It is typically used to classify data either in supervised or unsupervised manner. It is a rapid classification technique that is less susceptible to over-fitting data and can be run in a bootstrap fashion (Dinsdale et al. 2013). In addition, the random forest provides a measure of the importance of each variable that can be used in other analyses. There are several approaches that work in conjunction with random forests to estimates the importance of variables in separating the data into groups. One uses the mean decrease in accuracy that a variable causes is determined during the OOB (out-of-bag) error calculation phase. The values of a particular variable are randomly permuted among the set of OOB metagenomes. Then the OOB error is computed again. The more the accuracy of the random forest decreases due to the permutation of values of this variable, the more important the variable is deemed. The mean decrease in Gini is a measure of how a variable contributes to the homogeneity of nodes and leaves in the Random Forest (Dinsdale et al. 2013). Let p_{mgi} be the proportion of samples of group g_i in node m. Let g_c be the most plural group in node m. The Gini index of node ${}^{m}G_{m}$ is defined in the following equation (i)

$$G_m = \sum_{i \in g} p^{2}_{mgi} \quad \text{-------------} \ (1)$$

The Gini index is a measure of the purity of the node, with smaller values indicating a purer node and thus a lesser likelihood of misclassification (Brieiman et al. 2001). Tree generating algorithms may use this index as their likelihood to pick which variable to split on. Each time a particular variable is used to split a node, the Gini indexes for the child nodes are calculated and compared to that of the original node. When node m is split into m_r and m_l, there is a probability p_{m_r} of samples going into the child node m_r and p_{m_l} of going into m_l. The decrease (Brieiman et al. 2001) in Gini is defined in Equation (2)

$$D_m = G_m - p_{m_r} \, G_{mr} - p_{m_l} \, G_{p_l} \quad \text{-----------} \ (2)$$

The calculated decrease is added to the mean decrease Gini for the splitting variable and normalized at the end. The greater the mean decrease Gini of a variable, the purer the nodes splitting.

Each time a particular variable is used to split a node, the Gini coefficients for the child nodes are calculated and compared to that of the original node. The Gini coefficient is a measure of homogeneity from 0 (homogenous) to 1 (heterogeneous). The decreases in Gini are summed for each variable and normalized at the end of the calculation. Variables that split nodes into nodes with higher purity have a higher decrease in Gini coefficient.

Multidimensional scaling

Multidimensional scaling is a data visualization technique that directly scales objects based on either similarity or dissimilarity matrices (Quinn and Keough 2002). MDS takes for its input an n × n dissimilarity matrix S for n metagenomes, constructed by some other statistical technique, such as random forest. Then the algorithm looks for an embedding of the data points into some lower dimensional space that preserves the dissimilarity distances as much as possible. This embedding can then be plotted to visualize the clusters and their distances.

Linear discriminant analysis

Linear discriminant analysis is a robust supervised statistical technique that aims to separate the data into groups based on hyper planes and describe the differences between groups by a linear classification criterion that identifies decision boundaries between groups (Fisher 1936). Let X be a dataset with defined groups 1.........n. For each group j, there exists a corresponding conditional distribution describe in equation (3).

$$X(j) \,\square\, G(i) = j - f_j \text{ -------------------- (3)}$$

Furthermore, let π_j represent the proportion of X that is contained in group j. To perform a LDA on X, we assume that each f_j is normally distributed with an equal covariance matrix Σ, but with possibly different means μ_j. Using maximum likelihood estimation theory, the linear discriminant functions can be derived in equation (4).

$$g_j(x) = \log(\pi_j) + x\sum{}^{-1}\mu_j{}^{T} - \frac{1}{2}\mu_j\sum{}^{-1}\mu_j{}^{T} \text{ ------------ (4)}$$

These g_j's from (4) are our classifying functions. Since for a point x we sought to maximize $\pi_j f_j$, our classification criterion is

assign x to group j if $g_j(x)>g_k(x)$ for all $k\neq j$

With the classification criterion, decision boundaries between groups can be found. The decision boundaries are where the discriminant functions intersect. That is, the decision boundary between groups j and k is $\{x:g_j(x) = g_k(x)\}$. Therefore, the linear discriminant functions split the data space into regions. Each region corresponds to a specific group and the decision boundaries separate the regions.

Statistical software

The statistical and graphical methods discussed here are implemented using open source Statistical Language Programming R 3.2.2 (www.r-project.org).

Results and Discussion

RF generates a measure of the importance of each variable calculated by either the mean decrease in accuracy or the mean decrease in the Gini. These two values indicate which variables contributed the most to generating strong trees and can be used in MDS and LDA analyses. A subset of the data and variables is used to generate the trees and thus the approach can predict the environment to which a metagenome belongs. For both Accuracy and Gini in Fig.1 and Table 1, the photosynthesis got highest position with score 70.20 and 16.07 respectively as well as the phage groups with second highest score 61.31 and 12.14 for both procedures were the most important response variables in separating the datasets, and in the both cases a break occurred between these two variables and the remaining variables, suggesting that just these two measures could be used to grossly classify the metagenomes. Eight variables with highest Mean Decrease Accuracy and Mean Decrease Gini score were thus chosen for the following MDS and LDA analyses.

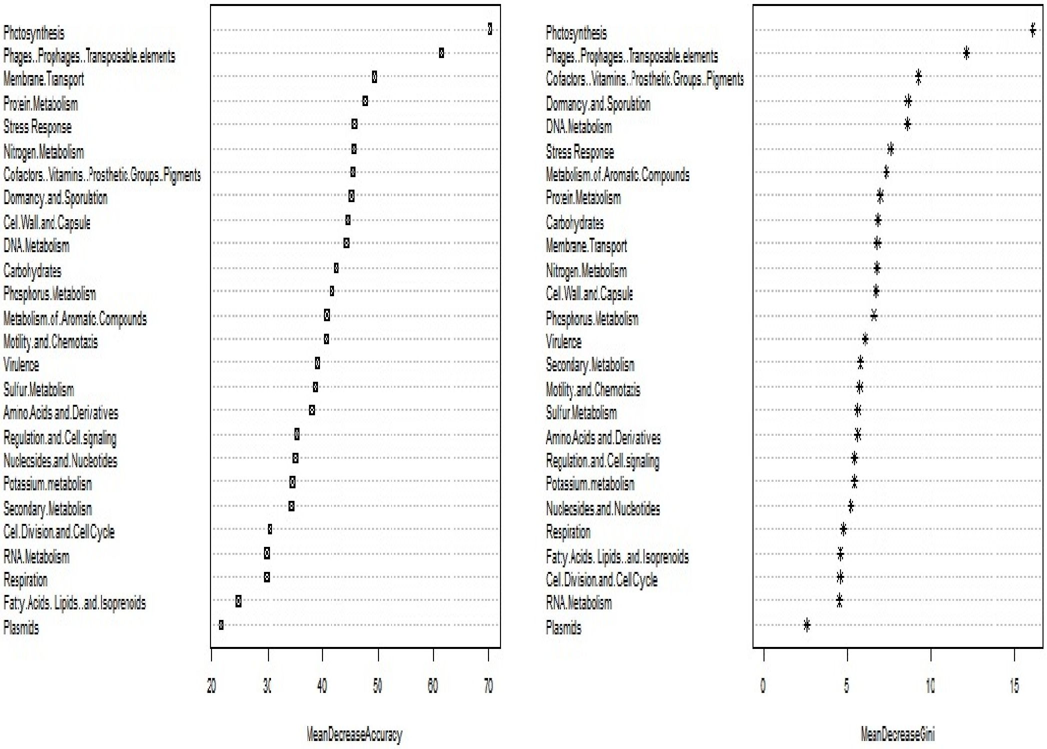

Fig. 1. Variable importance determined by random forest analysis using mean decrease in Accuracy and Gini. The plot outcome measures ranks of 26 the metabolic functions correspond to each symbols in both plots are placed according to their importance score in descending order.

Table 1. Variable importance measure with corresponding score.

Metabolic variables	Mean decrease in accuracy	Mean decrease in Gini	Metabolic variables	Mean decrease in accuracy	Mean decrease in Gini
Amino acids and derivatives	38.11	5.62	Phages, Prophages, and Transposable Elements	61.31	12.14
Carbohydrates	42.49	6.83	Phosphorus Metabolism	41.59	6.58
Cell Division and cell cycle	30.49	4.55	Photosynthesis	70.20	16.07
Cell wall and capsule	44.73	6.72	Plasmids	21.66	2.52
Cofactors, vitamins, Prosthetic groups, and pigments	45.49	9.2	Potassium Metabolism	34.54	5.40
DNA metabolism	44.44	8.58	Protein Metabolism	47.65	6.97
Dormancy and sporulation	45.29	8.61	Regulation and Cell Signaling	35.28	5.41
Fatty acids, lipids, and isoprenoids	24.84	4.57	Respiration	29.81	4.75
Membrane Transport	49.29	6.80	RNA Metabolism	29.82	4.52
Metabolism of Aromatic Compounds	40.68	7.34	Secondary Metabolism	34.40	5.76
Motility and Chemotaxis	40.56	5.73	Stress Response	45.79	7.60
Nitrogen Metabolism	45.68	6.76	Sulfur Metabolism	38.76	5.61
Nucleosides and Nucleotides	35.08	5.19	Virulence	39.11	6.05

MDS projects the proximity measures of the metagenomes as determined by RF to a lower-dimensional space (e.g., 2-dimensional space for plotting on xy-axis). For the RF, the similarity was measured as the number of times two metagenomes appeared on the same leaf in the trees (proximity), and is represented by the distance between two samples on the MDS plot. The MDS plots have been shown in Fig. 2 with the 10 predefined environments. In this analysis, the visualization highlights the separation of the microbes from human/animal hosts from other samples along the first dimension and the separation of the aquatic and mat communities along the second dimension.

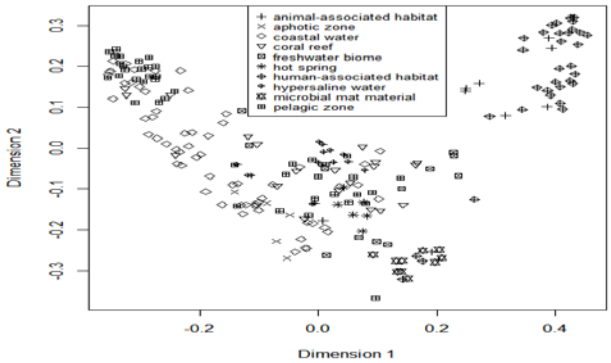

Fig. 2. Multiple dimensional scale plots of the distances calculated from unsupervised random forest. The distances are the number of times the samples appear on the same leaf of the tree, and the MDS has scaled them so that they plot projects those distances into two dimensions. Plotted by the original environments the sample came from.

In Fig. 3 the LDA overall 27 metabolic variables separated the data and showed that the human and terrestrial associated animal metagenomes separated from a cluster consisting of all of the aquatic samples except the hyper-saline community. The mat samples separated distinctly from other cluster

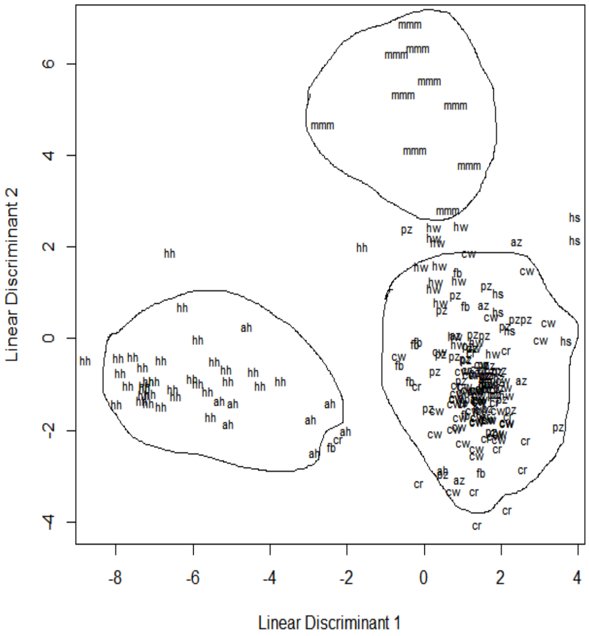

Fig. 3. Linear discriminate analysis showing the position of the metagenomes in two-dimensional space from the 10 environments.

For more detail and effective statistical analyses in terms of classification of the metagenomes correspond to their environments from which they belonged and sequenced, the robust statistical approaches should be employed. DNA sequence count of metagenomes may however suffer from the presence of extreme values. This kind of characteristics of the data increases the misclassification error rate and as a result provides low accuracy and precision of the statistical analyses. The robust or noble statistical or classification techniques will simultaneously deal the fact of presence of extreme as well as missing values in the DNA count dataset of metagenomes and provide decent appropriate explanatory and conclusive results.

Conclusion

The analyses separated the microbial samples into three broad groups: the human and animal associated samples, the microbial mats and the aquatic samples. The RF technique showed that phage activity is a major separator of host-associated microbial communities and free-living, suggesting that the phages are playing different ecological roles within each environment. The MDS and LDA techniques suggest that mat communities separated from both the animal associated metagenomes and the aquatic samples by the vitamin and cofactor metabolism, suggesting a role for secondary metabolism associated with growth in extreme environments. The dominant metabolic feature that separated the aquatic samples was photosynthesis. The marine environment categories of open ocean, coastal waters, coral reef and deep oceans share many metabolic features and therefore these metagenomes were placed into categories different than their a priori group assignment. This suggests subtle variation in metabolic processes that are occurring in the microbial communities from each environment that should be investigated in the future.

Acknowledgement

Authors would like to acknowledge to Bangladesh Agricultural Research Institute (BARI) for granting leave and Bioinformatics Lab., Department of Statistics, University of Rajshahi, Bangladesh for financial support for this study.

References

Angly F, Felts B, Breibart M, Salamon P, Edwards RA and Carlson CA (2006). The marine viromes of four oceanic regions. PLoS Biology 4: e368.

Aziz RK, Bartlets D, Best AA, Dejongh M, Disz T and Edwards RA (2008).The RAST Server: rapid annotations using subsystem technology, BMC Genomics 9: 75.

Breitbart M, Hoare A, Nitti A, Siefert J, Haynes M and Dinsdale E (2009). Metagenomic and stable isotopic analyses of modern freshwater microbialies in Cuatro CiEnegas, Mexico, Environmental Microbiology 11:16-34.

Brieman L (2001). Random Forests, Machine Learning 45: 5-32.

Brulc JM, Antonopoulos DA, Miller MEB, Wilson MK, Yannarell AC and Dinsdale EA (2009). Gene-enteric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases, Proceedings of National Academy of Sciences USA 106: 1948-1953.

Dinsdale EA, Edwards RA, Bailey BA, Tuba I and Akhter A (2013). Multivariate analysis of functional metagenomes, Frontiers in Genetics 4:1-24.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M and Brulc JM (2008). Functional metagenomic profiling of nine biomes. Nature: LETTERS, International Journal of Science 452: 628-629.

Fisher RA (1936). The use of multiple measurements in taxonomic problems, Annals of Eugenics 7: 179-188.

Huson DH, Auch AF, Qi J and Schuster SC (2009). MEGAN analysis of metagenomic data, Genome Research 17: 377-386.

Kurokawa K, Ttoh T, Kuwahara T, Oshima K, Toh H and Toyoda A (2008). Comparative metagenomics revealed commonly enriched gen sets in human gut microbiomes, DNA Research 14: 169-181.

Quinn GP and Keough MJ (2002). Experimental Design and Data Analysis for Biologists, Cambridge University Press.

Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K and Chang HW (2005). Comparative metagenomics of microbial communities, Science Microbiome 308: 544-557.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R and Gordon JI (2007). The human microbiome project. Nature: Insight Feature 449: 804-810.

Wooley JC, Godzik A and Friedberg I (2010). A primer on metagenomics, PLoS Computational Biology 6: e1000667.

Wogley L, Edwards RA, Rodriguez-Brito B, Liu H and Rohwer F (2007). Metagenomic analysis of the microbial community associated with the coral porties astreoides, Environmental Microbiology 9: 2707-2719.