Microbial Abundance Analysis and Phylogenetic Adoption in Functional Metagenomics

Jyotsna Talreja Wassan, Haiying Wang, Fiona Browne, Huiru Zheng*
School of Computing & Mathematics
Ulster University
Northern Ireland, United Kingdom

Wassan-JT@email.ulster.ac.uk, {hy. wang, f. browne, h.zheng*}@ulster.ac.uk

***Abstract*—Metagenomics is an unobtrusive science of studying uncultivated microbes sampled directly from an environment, e.g. soil, ocean, air, human body, or animals, etc. Functional metagenomics particularly deals with linking microbes to environmental derivations, such as classifying the role of human gut microbiome into a diseased or non-diseased state. Ongoing research in this area includes analyzing the structure of microbial communities, and relate it to functional analysis. We present an integrative experimental framework for functional metagenomics, including data driven (abundance count of microbial species) and knowledge driven (phylogenetic tree structure) contexts. Our related experiments, indicate that i) feature selection improves the performance of classifying human microbiome samples, ii) the classification of human microbiome remains a challenging problem while incorporating phylogenetic structures. For example, our best accuracy attained on the *Costello body site (CBH)* dataset with forehead and external ear as body sites, is *89.13 %* with a non-phylogenetic model, and *78.26 %* with a phylogenetic model. This forms a potential research direction of further exploration of space for incorporating phylogeny in microbial analysis and hence developing integrative computational models for deriving functional phenotypes, based on metagenomic sequencing data.**

Keywords—Metagenomics, Phylogeny, Classification, Next Generation Sequencing (NGS), Operational Taxonomical Units (OTUs), Metagenomes, Machine Learning (ML)

I. INTRODUCTION

Rapid advancement in Next Generation sequencing (NGS) techniques has expedited metagenomic analysis by generating DNA (whole genome or marker gene) sequences for significant environmental derivations [1]. The metagenomic pipelines like QIMME [2], Mothur [3], CloVR [4], etc., analyse the DNA sequences and convert it to abundance count matrix of Operational Taxonomic Units (OTUs), which represent binned metagenomic sequences at some similarity threshold (~97%) [5]. The OTUs are related by phylogeny, i.e., level of taxonomies from Phylum to Genus. The other outputs of the pipelines may range from phylogenetic tree, heat-maps to dissimilarity matrices like weighted or unweighted UniFrac matrix etc., which are further useful in downstream analysis of metagenomes [6]. Understanding the structural context of microbiome and linking it to functional roles is captivating research community. For example, the Human Microbiome Project (HMP) has paved path for studying human microbial samples from various body sites to detect various chronic human diseases, such as diabetes (Type 1 and Type 2), Inflammatory bowel disease (IBD Crohn's Disease), obesity (obese, lean, overweight), and cancer etc. [7,8]. Not just humans, but the field is impacting bio-sciences in general. In the recent years, the field has gained importance due to emergence of projects such as the HMP [9], Earth Microbiome [10], American Gut [11] and CAMERA [12]. However, it is a challenging research area due to key characteristics of metagenomic data, which are being plethoric, high dimensional, highly diverse, and sparse [13]. The metagenomic pipelines are generating biologically rich and computationally intensive datasets. Machine Leaning (ML) models [14-16] have been used in past for meta-analysis of such datasets but further extension, in terms of including phylogenetic information and OTU feature selection procedures, is likely to accelerate the learning.

In this paper, we present an integrative experimental framework for functional analysis of metagenomic datasets, based on: i) classification of raw OTU abundance count of taxas, and ii) linking phylogeny to raw-abundance analysis for determining environmental derivations using ML models. The framework serves as a potential road-map of possible experimental pathways leading to microbiome analysis. Along the pathways, we evaluated the computational models for metagenomic predictions. The current paper builds itself upon intriguing findings towards comparative evaluation of existing approaches involving identification of refined subset of independent OTU features for functional analysis and phylogeny driven prediction of metagenomic profiles. The remainder of the paper is organized as follows: related work is discussed in Section II. An integrative framework for functional metagenomics in the current study is described in Section III. The framework suggests that better performance is attainable with feature selection strategies over the raw OTU abundance count. Section IV provides materials, methods, and results. Finally, Section V provides the conclusion and future work.

II. RELATED WORK

The section presents key highlights on functional metagenomics, modelling of ML involved and adoption of phylogenetic context based on predominant notions from literature.

A. Functional Metagenomics

Functional analysis of metagenomic environments is a three-step process as listed below:

i. Input: A set of metagenomics sequences binned to OTU abundance count matrix, X (Eq. (1)), with m metagenomic samples and n OTUs; and set of functional labels Y.

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots\, x_{i,j} & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} \tag{1}$$

ii. A computational model that works on the input matrix X by providing a functional mapping from the row of X (representing a sample) to a functional label y ∈ Y

iii. Output: labelled sequences.

The main objective of this task is to predict the roles of microbial genes in functional derivations [5]. This task maps to the supervised classification in ML. The extent of metagenomic functional prediction tends to cater the following research avenues:

- Which OTUs determine characteristics of functional environment?
- What is the percentage of microbiomes with a given functional role (such as in a sample site or across samples)?
- How relationships and interactions between OTUs impact environmental derivations?
- How variations in composition of microbiome affect metagenomic analysis?

Cho et al. [7] discussed about metagenomic links to important functions in human health and disease. Ongoing research on varied phenotype hosts, reflects that the microbe-host interactions play important role in determining the influence on environmental condition being studied [10]. Nonetheless, functional analysis is accompanied by sparsity, diversity, and high variability in microbiome data. Taylor et al. [17] discussed various challenges and applications of functional assignment of metagenomic data. In the current study, we propose an experimental approach for effective functional characterization of microbiomes.

B. ML Methods in Microbial Biology

Supervised classification has been used to classify microbiota using OTU abundance count data. Knights. et al [14] performed supervised classification of human microbiomes using random forest (RF), nearest shrunken centroids (NSC), elastic net (ENet), support vector machines (SVM) with filters of bi-normal separation (BSS) and recursive backward feature elimination (RFE) over five benchmark datasets. Statnikov et al. [15] presented a comprehensive study of classifying human microbiomes using 18 classification methods involving RF, SVM, logistic regression (LR) and *K*-Nearest Neighbor (*k*-NN) with various parameter tunings and feature selection strategies of backward elimination with RF and RFE with SVM. Yang. et al. [18] classified the metagenomic samples from soil using SVM and *k*-NN classifiers.

The studies indicate that RF model yields the best performance in microbial studies [14, 15]. Also, SVM in concert with parameter tunings and RFE, has potential to provide satisfactory results. Wingfield. et al. [19] recently devised a hybrid classifier consisting of SVM with RFE and NN for metagenomic data analysis for characterizing Pediatrics Inflammatory Bowel Disease (IBD) in humans. In general, RF, SVM, linear and logistic models with Lasso & ENet, have been extensively applied in computational biology [14, 15, 18, 19]. However, the performance of classifier highly depends on the OTU features used in classification task. Hence, in current work, we intend to focus on selecting OTU features that entail better models and predictions over metagenomes.

C. Adoption of Phylogeny in Microbiome Functional Analysis

Phylogenetic context covers relatedness among microbial species in an ecological environment. In recent years, there has been advances in computational approaches for inferring phylogenetic relationships, via building phylogenetic tree structures and diversity indices such as α-diversity (phylogenetic diversity), β-diversity (UniFrac), etc. [20, 21]. Recent studies provide evidence of using phylogeny in the functional microbial analysis. Tanaseichuk et al. [22] proposed a novel supervised classification model, MetaPhyl, based on the multinomial Logistic Regression (LR) model with a tree-guided penalty function over the microbial features encoded in a phylogenetic tree, and presented an efficient optimization algorithm to learn the model regularization parameters. The model considered leaf level OTUs and lacked in considering taxas at multiple levels of the tree. Langille et.al. [23], devised *PICRUSt*, a phylogeny driven computational approach to predict the functional composition of a microbial communities using marker genes and database of reference genomes. Albanese et.al. [24] proposed an algorithm *PhyloRelief* to detect relevant microbial-taxa identification by applying the relief based strategy of feature ranking in phylogenetic context. Recently, Silverman et.al [25] presented *PhILR* transform combining statistical iso-metric log transform and phylogenetic context to analyse microbiomes. It is covetable to have deeper insights into phylogenetic adoption in metagenomic analysis. The integration of biological domain knowledge of related taxonomy with raw OTU abundance count data in microbiome analysis is an emerging arena for potential research.

III. The Proposed Framework

This section discusses the proposed integrative experimental framework for functional analysis in metagenomics, based on OTU count values of abundance and the phylogenetic context. Fig.1 depicts the proposed framework. The stages of the framework are described below.

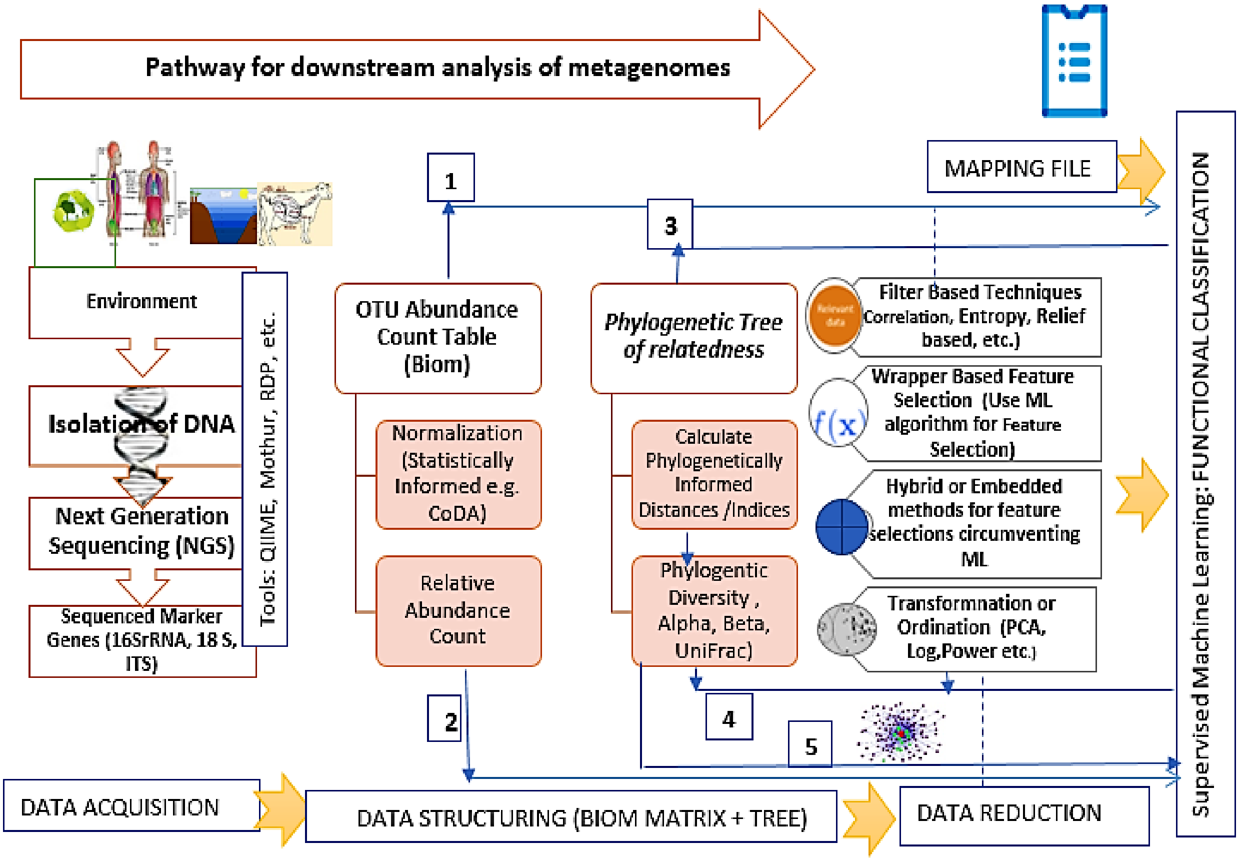
A. Data Acquisition

Next Generation sequencing techniques (Illumina /Solexa sequencing, Roche 454 sequencing/Ion torrent: Proton / PGM sequencing/ SOLiD sequencing) allow us to sequence extracted environmental DNA/ RNA at an unprecedented pace [1]. These techniques also aid in capturing marker genes (e.g. 16SrRNA) for metagenomic analysis. Use of metagenomic pipelines (e.g. QIIME/Mothur/CloVR etc.) over NGS obtained sequences, assist in achieving dominant genomes/taxas/OTUs with phylogenetic affiliations [2-4].

B. Data Structuring and Analysis Pathways

The output of above step is OTU abundance count table (BIOM) and a phylogenetic tree depicting relations between various OTUs. These structures serve as inputs for analysis under five possible pathways (Fig. 1) as listed below:

1. Applying ML (feature selection + classification) over raw OTU abundance count (pathway 1).
2. Applying normalization over OTU datasets to achieve relative OTU abundance count, which aids in variance stabilization amongst OTUs before application of ML models of feature selection and classification (pathway 2).
3. Including phylogenetic information from tree, in supervised ML algorithm, e.g. incorporating phylogeny in tuning parameters such as optimization / penalty/ regularization of classification ML algorithm or developing ML models for mining relationships between OTUs (based on weights on branches of phylogenetic tree), serving as most predictable in functional response to OTUs (pathway 3).
4. Calculating phylogenetically informed distances (calculated from tree) and using the attained values to geometrically transform or reduce OTU space before applying classification (pathway 4).
5. Developing association detection algorithm circumventing interactions between OTUs and thereafter using the attained output in predicting functional roles (pathway 5).



Pathway for downstream analysis of metagenomes
Environment
Isolation of DNA
Next Generation Sequencing (NGS)
Sequenced Marker Genes (16SrRNA, 18 S, ITS)
Tools: QIIME, Mothur, RDP, etc.
1
OTU Abundance Count Table (Biom)
Normalization (Statistically Informed e.g. CoDA)
Relative Abundance Count
2
3
Phylogenetic Tree of relatedness
Calculate Phylogenetically Informed Distances /Indices
Phylogentic Diversity , Alpha, Beta, UniFrac)
4
5
MAPPING FILE
Relevant data
Filter Based Techniques Correlation, Entropy, Relief based, etc.)
f(x)
Wrapper Based Feature Selection (Use ML algorithm for Feature Selection)
Hybrid or Embedded methods for feature selections circumventing ML
Transformation or Ordination (PCA, Log,Power etc.)
Supervised Machine Learning: FUNCTIONAL CLASSIFICATION
DATA ACQUISITION
DATA STRUCTURING (BIOM MATRIX + TREE)
DATA REDUCTION

Fig.1. A generalized experimental framework for functional metagenomics (1: pathway 1; 2: pathway 2; 3: pathway 3; 4: pathway 4, 5: pathway 5)

C. Data Reduction

Predictive models for distinguishing metagenomic functional roles depend on key microbial features and their associated roles. Selecting relevant OTUs, by applying feature selection (dimensionality reduction) strategies over high dimensional feature space potentially enhances the performance of predictive model in diverse datasets. The multitude of feature selection strategies are available [26]. However, the two dominant strategies are: i) filter based approach, and ii) wrapper based approach [26]. Filter methods are fast, based on statistical approaches over general data properties (such as feature rank and correlation with class), but are independent from classification model. On the other hand, wrapper methods determine OTU features by using the classification model that measures the relevance of each feature via evaluating feature-subsets over cross validations. The space search direction for feature selections could be forward, backward, or recursive elimination. Hybrid methods tend to combine the properties of filter and wrapper approaches and embedded methods serve as a model that has its own feature selection criteria in itself [26], [27]. Feature extraction is another way to characterize dimensionality reduction based on a mathematical transform. Principle Coordinate analysis (PCoA), Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), compositional transform like log-transform, power transform etc., are useful in functional metagenomics [26].

D. Final Step

This step incorporates application of ML supervised classification models over processed metagenomic datasets using the mapping file containing functional annotations. There exists multitude of ML models for classification [27]. The predominant ML methods chosen in current study are namely, Logistic Regression (LR), Random Forest (RF). Support Vector Machine (SVM) and Neural Networks (MLP), because of their successful application to cattle and human microbiome in our previous study [28]. LR [27], [28] considers the relationship between the class dependent variable (i.e. microbiome functional class) and one or more independent features (i.e. OTUs) by estimating probabilities based on logistic function (Eq. (2)).

$$P\,(Y = 1|X) = \mathrm{f}\,(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}}, \tag{2}$$

where t is a linear combination of features x_{i}'s, and is equivalent to $b_0 + b_1x_1 + \cdots.\ b_nx_n$; and b_i represents the numerical coefficients of estimation where i ranges from 1 to n features. RF [28] works by constructing an ensemble of decision tree-structured classifiers with bagging. Several decision trees are trained with random bootstrap samples from the original microbiome space to provide classification results by means of voting or averaging the results over different trees. SVM [28], tends to project n OTU data points into an n-1-dimensional space, in which the functional classes are linearly separable, and to identify the maximum-margin hyperplane that maximizes the distance between the classes. MLP [28], is inspired from the working of human brain learning layers. The input OTU feature vectors are forwarded to sequenced layers of neurons in combination with associated weight thresholds, which drives the ability to perform classification at output layer on some excitation threshold. The proposed workflow projects to answer the following two main research questions in future for metagenomic functional analysis:

1. Which OTUs are relevant and serve as most predictable in response of their abundance to functional roles?
2. Which OTUs correlate and what role does the association between OTUs play, with varied phylogenetic depth in analysis of metagenomes?

This framework proposes an experimental set-up based on data driven and knowledge driven context in functional analysis of metagenomes.

IV. Material, Methods & Results

This section depicts preliminary results achieved in functional classification over metagenomic Use Case datasets, by applying the proposed framework section III.

A. Datasets under Study

The experiments were based on applying ML models on two publicly available datasets [24]: - i) Inflammatory bowel disease (IBD) dataset with 75 samples and 199 OTUs, to distinguish between IBD(diseased) state from non-IBD state based on microbial communities' present in human fecal samples [29], and ii) human microbiome Costello Body Habitats (CBH) dataset with 46 samples and 283 OTUs, a benchmark that includes samples from microbial communities' present in forehead and external nose as human body sites [30].

B. Models Used for Functional Classification

We used 16 non-phylogenetic and 2 phylogenetic models to classify metagenomes. The non-phylogenetic were based on ML models: LR, RF, SVM and MLP (as were discussed in section III). The configuring parameters of ML models aided in attaining high performance in this study, are listed in Table I. We emphasize that all non-phylogenetic models were applied using 10 folds' cross validation (10-fCV), to entail better performance.

TABLE I. ML MODELS WITH PARAMETER CONFIGURATIONS

ML Model	Parameter Configurations
LR	Regularization: Ridge of 1.0E-8
RF	Maximum Depth of Tree: 6, Number of Iterations: 100
SVM	Kernel: PolyKernel with Exponent 1 Calibrator: Logistic Function Complexity parameter: 1
MLP	Hidden Layers: 2 Learning Rate: 0.3 Momentum: 0.2

We circumvent the problem of high-dimensional nature of OTUs by using following features selection strategies: - the wrapper based method using logistic regression and random forest as ML models (WB-LR & WB-RF), and filter method based on correlation of OTUs with functional class (CFS). We also experimented with two phylogenetic driven models from literature, namely MetaPhyl and PhILR phylogenetic transform [22, 25]. MetaPhyl is an executable, C++ binary available publicly [1] and uses multinomial logistic regression model,

[1] http:// alumni.cs.ucr.edu/~tanaseio/metaphy.htm

trained using penalty function, gaining insights from OTU relationships encoded in phylogenetic tree for optimization [22]. PhILR package[2] ("philr") is available in R [25]. The aim of using PhILR is to transform original OTU datasets into an orthogonal unconstrained space. The transform produced balances (BT) representing the log-ratio of the geometric mean abundance of the two groups of OTU features descending from same parent in a phylogenetic tree [25]. LR was applied over the obtained OTU balances (BT) from the transform (PhILR +LR). The OTU datasets were pre-processed using "Phyloseq" package[2] in R. The datasets were normalized to handle the high volume of zero values present in the dataset. The models were evaluated on the overall prediction accuracy (Acc.) representing number of samples correctly classified; root squared mean error (RMSE), which represents concentration of data around the line of best fit during classification; precision (Pr.) which depicts percentage of retrieved instances that are relevant, and area under curve (ROC) measure, representing the performance of a classifier with variation in discrimination threshold [31].

C. Results and Discussion

The results of models as listed above in section B are shown in Tables II, IV. The results indicate that the process of feature selection with wrapper and filter methods, reduces the dimensionality of the raw OTU feature vectors whilst providing improved levels of predictive accuracy, precision, and ROC. It also reduced the RMSE in comparison to analysis over raw OTU abundance counts. We report average performance values in analytical results for brevity. The ML model, WB-LR+LR, over independent raw OTU abundances in IBD dataset, provided best result in terms of accuracy (=90.66) and ROC (=0.855) (Fig. 2). The overall results achieved over IBD dataset, are shown in Table II. We found that taxas of phylum *Firmicutes* and *Bacteroidetes* play major role in differentiating between subjects of non-IBD (healthy) and IBD (diseased), using the model over IBD dataset with 10 folds cross validation for classification (10-fCV). The predominant differentiating OTUs are listed in Table III. The method achieved higher performance of classification than phylogenetic methods of: - MetaPhyl and PhILR and over the established RF model in literature [14,15], for IBD Use Case. WB-LR+MLP also attained competent performance in this Use Case. On the other side, considering the phylogenetic pathway, MetaPhyl provided good accuracy of 89.33, over all OTUs.

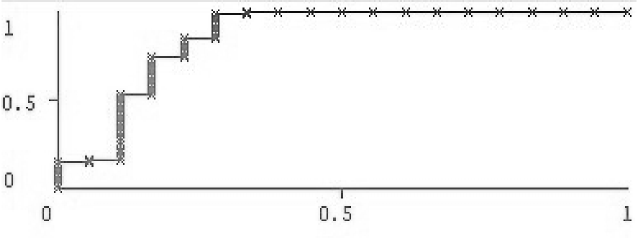


Fig. 2. The best ROC area under the curve value attained (using WR-LR+LR) for IBD subjects where X axis represents False Positive Rate & Y axis represents True Positive Rate

TABLE II. RESULTS OF FUNCTIONAL ANALYSIS OVER HUMAN MICROBIOME IBD DATASET (NF: NUMBER OF OTUS, BT: NUMBER OF BALANCES)

Dataset 1 (75X 199), # of classes 2	**ML**		**Performance**			
	Model with Test Mode: 10- f CV	***NF/BT***	***Acc. (%).***	***RMSE***	***Pr.***	***ROC***
No Feature Selection	LR	199	61.3	0.6074	0.647	0.547
	RF (depth 6)	199	77.3	0.3877	0.751	0.779
	SVM	199	65.3	0.5888	0.653	0.525
	MLP (layer=2)	199	66.6	0.5466	0.685	0.628
	CFS+ LR	12	77.3	0.4363	0.761	0.754
	CFS+RF	12	78.6	0.3822	0.767	0.815
	CFS+SVM	12	78.6	0.4619	0.764	0.632
	CFS+MLP	12	72	0.4726	0.715	0.719
Feature Selection	**WB-LR+LR**	**8**	**90.66**	**0.2972**	**0.905**	**0.855**
	WB-LR + RF	8	82.6	0.3864	0.820	0.761
	WB-LR+SVM	8	81.3	0.432	0.817	0.630
	WB-LR +MLP	**8**	**90.6**	**0.2938**	**0.917**	**0.818**
	WB-RF+LR	11	77.3	0.4205	0.754	0.701
	WB-RF +RF	**11**	**86.66**	**0.3474**	**0.863**	**0.835**
	WB-RF+SVM	11	73.33	0.5164	0.638	0.501
	WB-RF+MLP	11	73.3	0.4689	0.723	0.685
Phylogenetic Models	**MetaPhyl (C++)**	**199**	**89.33**	**-**	**0.796**	**-**
	PhILR + LR	4	73.36	0.4027	0.703	0.742

TABLE III. OTUS DIFFERENTIATING IBD & NON-IBD CASES

S.No.	**Predominant OTUs in IBD Classification**
1	k__Bacteria_p__Firmicutes_c__Clostridia_o__Clostridiales_f__Ruminococcaceae_
2	k__Bacteria_p__Bacteroidetes_c__Bacteroidia_o__Bacteroidales_f__Bacteroidaceae_g__Bacteroides
3	k__Bacteria_p__Bacteroidetes_c__Bacteroidia_o__Bacteroidales_f__Odoribacteraceae_g__Odoribacter_s__
4	k__Bacteria_p__Firmicutes_c__Clostridia_o__Clostridiales_f__Veillonellaceae_g__Megasphaera
5	k__Bacteria_p__Bacteroidetes_c__Bacteroidia_o__Bacteroidales_f__Porphyromonadaceae_g__Parabacteroides
6	k__Bacteria_p__Bacteroidetes_c__Bacteroidia_o__Bacteroidales_f__Bacteroidaceae_g__Bacteroides_s__plebeius
7	k__Bacteria_p__Firmicutes_c__Clostridia_o__Clostridiales_f_Tissierellacea
8	k__Bacteria_p__Firmicutes_c__Bacilli_o__Lactobacillales_f__Leuconostocaceae

The modelling with wrapper feature selection with WB-LR+LR, WB-LR+MLP and WB-LR+RF also provided best results in terms of performance in comparison to other models applied over OTU abundances in CBH dataset, (Table IV). The highest accuracy was attained by WB-LR+LR (=89.13 %) which is better than MetaPhyl (=78.26 %) and PhILR (=73.91%) models of phylogeny. The highest ROC (= 0.827), was achieved with WB-LR + RF in this Use Case (Fig. 3). We found that taxas of phylum *Firmicutes*, Proteobacteria and *Actinobacteria* differentiate best between body sites of forehead and external ear with 10 folds cross validation for classification (10-fCV). The predominant differentiating OTUs attained on basis of WB-LR feature selector, are listed in Table V. In this case, we also found that CFS with LR significantly outperformed ML models applied on raw OTU data and the two phylogenetic models.

[2] https://bioconductor.org/packages

TABLE IV. RESULTS OF FUNCTIONAL ANALYSIS OVER HUMAN MICROBIOME CBH DATASET (BODY SITES: FOREHEAD & EXTERNAL EAR) (NF: NUMBER OF OTUS, BT: NUMBER OF BALANCES)

Dataset2 (47X 283), # of classes 2	**ML**	**Test Mode: 10- f CV**	**Performance**			
	Model	***NF / BT***	***Acc. (%).***	***RMSE***	***Pr.***	***ROC***
No Feature Selection	LR	283	65.21	0.58	0.638	0.596
	RF (depth 6)	283	67.39	0.465	0.587	0.718
	SVM	283	63.04	0.6079	0.638	0.565
	MLP (layers=2)	283	69.5	0.5175	0.684	0.706
Feature Selection	**CFS+ LR**	**3**	**84.7**	**0.3875**	**0.875**	**0.665**
	CFS+RF	3	78.2	0.3953	0.777	0.816
	CFS+SVM	3	76.1	0.489	0.822	0.602
	CFS+MLP	3	82.6	0.3875	0.861	0.698
	WB-LR+LR	**7**	**89.13**	**0.3538**	**0.906**	**0.767**
	WB-LF +RF	7	73.91	0.4324	0.724	0.711
	WB-LR+SVM	7	78.26	0.4663	0.834	0.636
	WB-LR +MLP	**7**	**80.43**	**0.373**	**0.814**	**0.704**
	WB-RF+LR	7	69.56	0.4737	0.665	0.571
	WB-RF+RF	**7**	**84.7**	**0.3743**	**0.853**	**0.827**
	WB-RF+SVM	7	69.56	0.5517	0.642	0.519
	WB-RF+MLP	7	69.56	0.4457	0.665	0.674
Phylogenetic Models	**MetaPhyl(C++)**	**283**	**78.26**	**-**	**0.642**	**-**
	PhILR +LR	7	73.91	0.4563	0.724	0.688

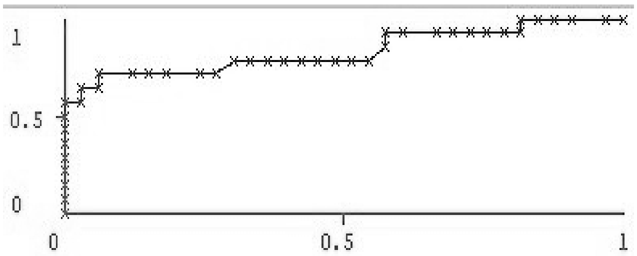


Fig. 3. The best ROC area under the curve value (using WR-LR+RF) attained for Costello Body Sites (External ear) where X axis represents False Positive Rate, & Y axis True Positive Rate

The wrapper and filter based approaches in combination with RF attained higher performance in comparison to RF model used over raw OTU count in Use Cases.

Also, on comparing computational methods for classifying Human microbiomes, over both the Use Cases, wrapper based on LR, achieved best performance in terms of accuracy. The phylogenetic models and filter based methods provide better performance than ML models applied over raw abundance count. It reflects that the suitable combination of feature selection of OTU feature space and a classification algorithm, plays important role in functional analysis of microbiomes.

TABLE V. OTUS DIFFERENTIATING BETWEEN COSTELLO BODY SITES

S.No.	**Predominant OTUs in Classification of human body sites (Forehead & External Ear)**
1	k__Bacteria_p__Proteobacteria_c__Gammaproteobacteria_o__Pseudomonadales_f__Moraxellaceae_g__Acinetobacter
2	k__Bacteria_p__Actinobacteria_c__Actinobacteria_o__Actinomycetales_f__Corynebacteriaceae_g__Corynebacterium
3	k__Bacteria_p__Proteobacteria_c__Alphaproteobacteria_o__Sphingomonadales_f__Sphingomonadaceae_g__Sphingobium
4	k__Bacteria_p__Firmicutes_c__Clostridia_o__Clostridiales_f__Tissierellaceae_g__Anaerococcus
5	k__Bacteria_p__Firmicutes_c__Bacilli_o__Bacillales
6	k__Bacteria_p__Proteobacteria_c__Alphaproteobacteria_o__Rhodobacterales_f__Rhodobacteraceae_g__Rubellimicrobium
7	k__Bacteria_p__Actinobacteria_c__Actinobacteria_o__Actinomycetales_f__Propionibacteriaceae_g__Propionibacterium_s__granulosum

V. CONCLUSION AND FUTURE WORK

In this paper, we characterized an experimental set-up for analyzing microbial communities in accordance to their sample types, based on their taxonomic profiles. An overarching goal of the proposed set-up is to classify microbial communities into their functional phenotypes We conducted an evaluation study to identify accurate ML models for classifying human microbiomes by following two primary directions: i) non-phylogenetic and ii) phylogenetic.

We followed pathway 1 of proposed framework (Fig.1) by applying feature selection strategies and classification models over multivariate metagenomic data, pathway 3 by using MetaPhyl [22] method and pathway 2 and 4 by using PhILR [25] method.

The results reveal that the computational performance (accuracy) does not always necessarily improve with phylogenetic models, which otherwise are expected to produce more significant biological results. It is being believed that phylogenetic complexity influences the downstream analysis in metagenomics. But we support that incorporating phylogeny neither precludes nor outperforms the computational performance in functional metagenomics analysis.

The results also indicate that the best combination of OTU feature selection and classifier to determine functional repertoire of human microbiome in current study is: - feature subset selection with wrapper methods (based on LR, RF learners), over OTUs, and supervised ML learners (LR, RF and MLP) as classifiers. This provides an improvement over RF which is the most popular non-phylogenetic technique for microbiome classification. These combinations also provide better degree of accuracy in comparison to the phylogenetic models used in analysis.

However, phylogeny driven models do offer improvement upon ML classifiers (LR, SVM, RF, and MLP) when applied over raw OTU data alone.

In future, we would like to build an integrative model based on phylogenetic structures to achieve better performance over metagenomic functional predictions by anticipating the use of following possible approaches:

- Applying compositional analysis (CoDA) techniques[3] and performing ML over normalized metagenomes (pathway 2 in proposed framework [Fig.1])
- Exploring the space further for in-cooperating phylogenetic context into ML models for microbial analysis to combine knowledge-based approaches with data driven methods for achieving computationally better and biologically significant functional predictions (pathway 3 & 4 in proposed framework [Fig.1]).
- Constructing network based models for studying microbial interactions to seek coverage over diversity in metagenomes by studying co-occurrence or co-abundance patterns (pathway 5 in proposed framework [Fig 1]).

We speculate that the incorporation of biological and structural context would provide more realistic and significant modelling for predicting functions in metagenomics.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. C. Schuster, “Next-generation sequencing transforms today’s biology,” *Nat. Methods*, vol. 5, no. 1, pp. 16–18, 2008.

[2] J. Kuczynski, J. Stombaugh, W. A. Walters, A. González, J. G. Caporaso, and R. Knight, “Using QIIME to analyze 16s rRNA gene sequences from microbial communities,” *Curr. Protoc. Microbiol.*, 2012.

[3] P. D. Schloss *et al.*, “Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities,” *Appl. Environ. Microbiol.*, vol. 75, no. 23, 2009.

[4] J. White, C. Arze, M. Matalka, T. C. Team, S. Angiuoli, and W. F. Fricke, “CloVR-16S: Phylogenetic microbial community composition analysis based on 16S ribosomal RNA amplicon sequencing – standard operating procedure, version1.0.,” *Nat. Preced.*, pp. 1–9, 2011.

[5] J. C. Wooley, A. Godzik, and I. Friedberg, “A primer on metagenomics,” *PLoS Computational Biology*, vol. 6, no. 2. 2010.

[6] D. Marco, *Metagenomics: Theory, Methods and Applications*, vol. 7, no. Suppl 5. 2010.

[7] I. Cho and M. J. Blaser, “The human microbiome: at the interface of health and disease.,” *Nat. Rev. Genet* , vol. 13, no. 4, pp. 260–270, 2012.

[8] D. McDonald, A. Birmingham, and R. Knight, “Context and the human microbiome.,” *Microbiome*, vol. 3 p. 52, 2015.

[9] The NIH HMP Working Group, “The NIH Human Microbiome Project,” *Genome Res.*, vol. 19, no. 12, pp. 2317–2323, 2009.

[10] J. A. Gilbert *et al.*, “The Earth Microbiome Project,” in *1st EMP meeting on sample selection and acquisition*, 2010.

[11] American Gut Project. http://americangut.org/about/. Accessed May 2017.

[12] R. Seshadri, S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier, “CAMERA: A community resource for metagenomics,” *PLoS Biology*, vol. 5, no. 3. pp. 0394–0397, 2007.

[13] H. Teeling and F. O. Glöckner, “Current opportunities and challenges in microbial metagenome analysis-A bioinformatic perspective,” *Brief. Bioinform.*, vol. 13, no. 6, pp. 728–742, 2012.

[14] Knights et al., 2010, “Supervised classification of human microbiota - FEMS Microbiology Reviews , Wiley Online Library.”

[15] A. Statnikov *et al.*, “A comprehensive evaluation of multicategory classification methods for microbiomic data.,” *Microbiome*, vol. 1, no. 1, p. 11, 2013

[16] H. Soueidan and M. Nikolski, “Machine learning for metagenomics: methods and tools," pp. 1–23, 2015.

[17] T. Prakash and T. D. Taylor, “Functional assignment of metagenomic data: Challenges and applications,” *Brief. Bioinform.*, vol. 13, no. 6, pp. 711–727, 2012.

[18] C. Yang *et al.*, “An ecoinformatics tool for microbial community studies: Supervised classification of Amplicon Length Heterogeneity (ALH) profiles of 16S rRNA,” *J. Microbiol. Methods*, vol. 65, 2006.

[19] B. Wingfield and S. Coleman, “A Metagenomic Hybrid Classifier for Paediatric Inflammatory Bowel Disease,” pp. 1083–1089, 2016.

[20] J. Dees, J. L. Momsen, J. Niemi, and L. Montplaisir, “Student interpretations of phylogenetic trees in an introductory biology course,” *CBE Life Sci. Educ.*, vol. 13, no. 4, pp. 666–676, 2014

[21] C. M. Tucker et al., “A guide to phylogenetic metrics for conservation, community ecology and macroecology,” Biol. Rev., vol. 92, pp. 698–715, 2016.

[22] O. Tanaseichuk, J. Borneman, and T. Jiang, “Phylogeny-based classification of microbial communities,” *Bioinformatics*, vol. 30, no. 4, pp. 449–456, 2014.

[23] M. G. I. Langille *et al.*, “Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.,” *Nat. Biotechnol.*, vol. 31, no. 9, pp. 814–21, 2013

[24] D. Albanese, C. De Filippo, D. Cavalieri, and C. Donati, “Explaining Diversity in Metagenomic Datasets by Phylogenetic-Based Feature Weighting,” *PLoS Comput. Biol.*, vol. 11,pp. 1–18, 2015.

[25] J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David, “A phylogenetic transform enhances analysis of compositional microbiota data,” *Elife*, vol. 6, pp. 1–20, 2017.

[26] Z. M. Hira and D. F. Gillies, “A review of feature selection and feature extraction methods applied on microarray data,” *Adv. Bioinformatics*, 2015

[27] Lee, J. Won, et al. "An extensive comparison of recent classification tools applied to microarray data.", Computational Statistics & Data Analysis ,48.4 , 2005.

[28] J.T. Wassan et al., “An Integrative Approach for the Functional Analysis of Metagenomic Studies,” unpublished.” 2017.

[29] P. Eliseo, et al. "Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease." *PloS one* 7.6, 2012.

[30] Costello, Elizabeth K., et al. "Bacterial community variation in human body habitats across space and time." *Science* 326.5960, 2009.

[31] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, 2009.

[3] http://www.compositionaldata.com