An Integrative Approach for the Functional Analysis of Metagenomic Studies

Jyotsna Talreja Wassan[1], Haiying Wang[1], Fiona Browne[1], Paul Wash[2], Brain Kelly[2], Cintia Palu[2,4], Nina Konstantinidou[2], Rainer Roehe[3], Richard Dewhurst[3], Huiru Zheng[1*]

[1] School of Computing and Mathematics, Ulster University, N. Ireland, United Kingdom
[2] NSilico Pvt. Ltd., Ireland
[3]Future Farming Systems, SRUC (Scotland's Rural College), Scotland, United Kingdom
[4]University College Cork
Corresponding author: h.zheng@ulster.ac.uk

Abstract. Metagenomics is one of the most prolific "omic" sciences in the context of biological research on environmental microbial communities. The studies related to metagenomics generate high-dimensional, sparse, complex, and biologically rich data-sets. In this research, we propose a framework which integrates omics-knowledge to identify suitable-reduced set of microbiomes features, for gaining insights into functional classification of the metagenomic sequences. The proposed approach has been applied to two Use Case studies, on: 1) cattle rumen microbiota samples, for differentiating nitrate and vegetable oil treated feed, for improving cattle performance, under MetaPlat H-2020 Project[1], and 2) human gut microbiota and classifying them in functionally annotated categories of leanness, obesity, or overweight. A high *Accuracy* of 97.5 % and *Area Under Curve* performance value (AUC) of 0.972 was achieved for classifying Bos *taurus*, cattle rumen microbiota data samples using Logistic Regression (LR) as classification model as well as feature selector in wrapper based strategy for Use Case 1 and 94.4 % *Accuracy* with AUC of 1.000, for Use Case 2 on human gut microbiota. In general, *LR classifier with Wrapper- LR learner (with ridge estimator) as feature selector*, proved to be most robust in analysis.

Keywords: Metagenomics, OTUs (Operational Taxonomic Units, Phylogeny, Machine Learning (ML), Classification

1 Introduction

Metagenomics involves the study of gene sequences of microorganisms derived directly from the natural environment such as air, water, human or animal body, and soil etc., following a culture-independent approach [1]. In the past few years, this field

[1]MetaPlat (http://www.metaplat.eu) is a 4-year project funded by European Horizon H2020-MSCA-RISE-2015

has gained prominence due to important projects such as the Human Microbiome Project(http://hmpdacc.org/), Earth Microbiome Project (http://www.earthmicro biome.org/) , and American Gut Project(http://americangut.org/), and due to unprecedented advances in low cost DNA isolation and sequencing strategies such as high speed throughput Next-Generation Sequencing (NGS) over the traditional Sanger approach [2, 3]. Several studies have shown relations between microbial diversity and host phenotypes. For example, human microbiome is related to various diseases such as diabetes (Type 1 and Type 2), Inflammatory bowel disease (IBDCrohn's Disease or Healthy), Obesity (Obese, lean, overweight), and cancer etc. [2, 4]. Belanche et.al. [5] recently studied the impact of supplementing Grass Hay with Vitamin A on rumen microbiome and its function. Roehe et al. [6] found that host genetics is shaping the rumen microbiome influencing methane production and feed conversion efficiency in cattle.

Metagenomic studies follow a typical metagenomic pipeline consisting of various stages including, gene sampling, sequencing, assembly, binning, taxonomic assignment, functional data analysis, and data sharing [7]. The binning of sequences generates Operational Taxonomic Units (OTUs)/taxas. OTU abundance count, relations between OTUs (phylogeny) and sample microbe-microbe interactions contribute effectively in analyzing metagenomic functional roles. Current computational challenges along the metagenomic pipeline concern data management, processing, and analysis of metagenomics datasets. These are due to key characteristics of metagenomic data, being massive, high dimensional, sparse, heterogeneous, incomplete, highly-skewed, and noisy [8, 9]. Emergence of NGS, has resulted in a gap between the pace of data generation and its analysis [3]. The variance in OTU abundance count also does not follow a normal distribution, and pose statistical challenges [9]. Considering these challenges, we propose an integrative approach, combining omics and data analytics to identify functional roles of metagenomic datasets. This is achieved by identifying a subset of OTU features which offer optimal predictive modelling built upon various Machine Learning (ML) classification algorithms. Selecting a subset of relevant OTU features for ML models is expected to entail improvement in performance.

2 Materials

The study involves analysis over two Use Case datasets; i) B. *taurus* (cow) rumen microbiota dataset and ii) human distal gut microbiome dataset. The B. *taurus* microbiota plays an important role in cattle productivity, health, and immunity. To investigate B. *taurus* gut microbiota in the context of these environmental traits, it's community composition was determined in 40 case samples provided by the MetaPlat project[1]. The data consist of 20 samples from an oil based treatment and 20 samples from a nitrate based treatment to reduce methane emissions. 5 OTU tables, with different taxonomic levels (Phylum to Genus) of classification, were generated in QIIME by NSilico (http://www.nsilico.com/). The tables consist of 27, 52, 101, 194 and 386 OTU feature vectors for phylum, class, order, family, and genus levels respectively. The dataset under consideration for second Use Case was obtained from a study on the human gut microbiome in obese and lean twins conducted by Turnbaugh et.al. [4]. To address the

factors related to obesity and genotypes, the study considered the microbiome of twin pairs and their mothers using 16s rRNA genes. The dataset consists of 18 microbiome samples, 756 OTU species and 3 classes (lean, obese, overweight) for analysis.

3 Methodology

This research was performed using NGS-16S genomic datasets listed in Section 2. OTU tables (Biological Observation Matrix), consisting of raw abundance counts, were obtained using the QIIME(http://qiime.org/) or CloVR-metagenomics pipelines (http://data.clovr.org/d/10/obese-lean-twin-gut-metagenome-output) [10]. The samples also associated meta-data describing their relationship with environmental traits. The OTU tables were pre-processed and transposed to fit to ML models. To maximize the performance of our experimental design, we followed an integrated workflow (as depicted in Fig.1.), focusing on two major steps: (1) selecting a suitable feature selection method and (2) selecting an appropriate learning classification functional model over selected features in Step (1), by evaluating its performance.

3.1 Feature Selection

Feature selection methods remove irrelevant and redundant features. The process primarily consists of two main steps: - i) feature subset search and ii) feature subset evaluation. We employed Best First Search (BFS) and Ranker's Method (RM) as feature search strategies [11], to the OTUs at various taxonomical levels. BFS is based on back-tracking the search path for finding OTU subsets till prominent results are attained whereas in RM, OTU features are ranked by their individual evaluations over selection metrics like associated weights, entropy, etc. on a user defined threshold. The features exceeding a threshold defined by user are selected for further analysis. The default value of threshold is set to -1.79769 in our analysis.

The next differentiating factor after subset search, is to evaluate the attained subsets for application of ML models. The evaluations are typically inspired from two categories: i) filter based techniques (FFS), in which function evaluates the worth of features by heuristics over general characteristics of OTU data or ii) wrapper based (WFS) techniques which evaluates the worth by using an embedded ML algorithm over OTUs [11]. The various filter techniques used for evaluating OTU data are Correlation-based feature selection (CFS), selecting OTU features that are highly correlated with the class but uncorrelated with each other; Info-gain based feature selection (IFS) which is driven from probabilistic modelling of nominal valued feature subsets; Principal Component Analysis (PCA) that transforms existing features in the subset to new features in lower dimensional feature space; and Relief based Evaluation (RB), evaluating the worth of OTU features by instance based learning [11]. Wrapper Based Filters evaluate attribute sets by estimating their accuracy using a learning scheme [11]. The related user defined parameters include the classifier model, the number of associated folds (set to 5), random seed (set to 1) and any associated threshold value (set to 0.01) in our current study. The classifier model supports a variety of algorithms from Naïve Bayes probabilistic

measure, Support Vector Machines, K-Nearest Neighbor, Logistic Regression to Random Forests, Boosting, etc. The most discriminative OTU feature selection has potential to reduce complexity of the potential ML model and increase the performance.

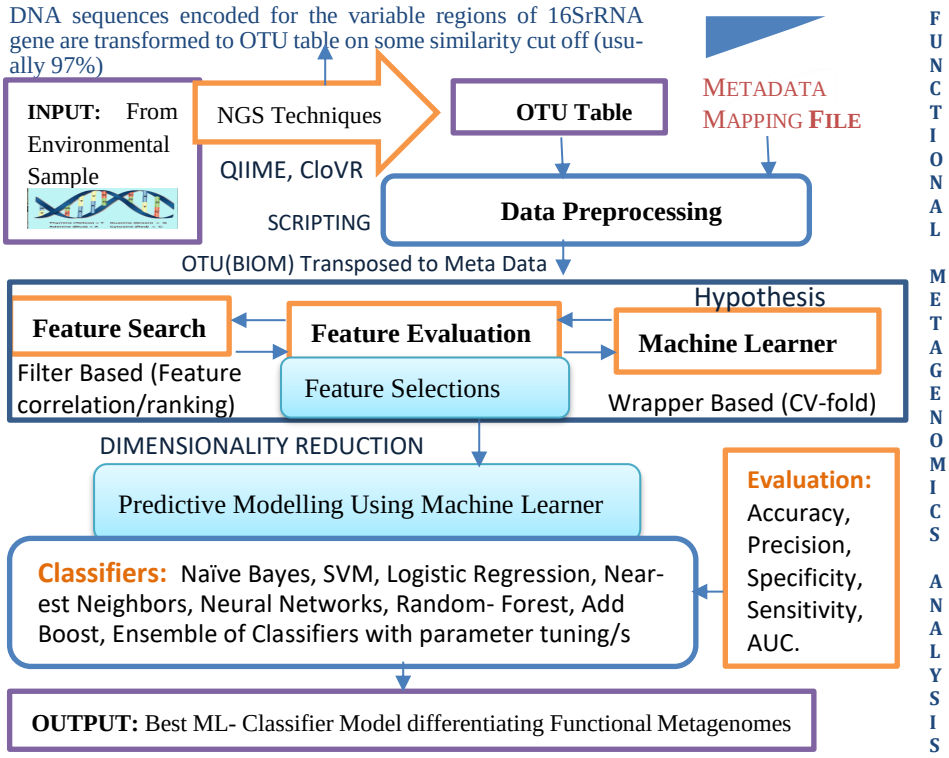

Fig. 1. An integrated approach for functional analysis in metagenomics

3.2 Learning Functional Models

A machine learning model works over the knowledge induced from the sample OTU data sub-sets attained in section 3.1. Applying ML models for categorizing the OTU features, into one of a pre-specified set of functional categories, is the key characteristic step of functional metagenomics. To identify the most suitable model for predicting functional metagenomes, various supervised ML classification algorithms were evaluated for their fitness in the prediction task against the selected OTU feature set. The range of classifiers applied are: *Naïve Bayes (NB)* with kernel estimator as false; *Neural Network (NN)*, with hidden layers as 01/02/ no=(features + classes)/2, random seed as 1-10, validation threshold as 20, model learning rate as 0.3, momentum as 0.2; *Random Forest (RF)* with maximum tree depth as 0-6; *Support Vector Machine (SVM)* with Poly-Kernel/RBF Kernel and c, seed parameters varying between 0-10; *Logistic Regression (LR)* with ridge estimation; *k-Nearest-Neighbor's classification (K-NN)* with linear nearest neighbor search and no: of neighbors as 1-5; *Adaptive Boosting (AdaB)*

and an *ensemble of classifiers*(Zero-R, NN, K-NN, LWL) [12,13]. To assess the performance of each prediction model, a 10-folds cross validation procedure was carried out in which OTU data is randomly split into k = 10, mutually exclusive subsets of equal size for overall assessment of classification. The performance assessment metrics used for evaluating classification models, in our study, are: Accuracy (Ac.), Precision (Pr.), Sensitivity (Se.), and Specificity (Sp.), Area under Curve (AUC-ROC) [13, 14].

4 Experiments and Results

Predictive modelling over the Use Cases supports holistic understanding of input data behavior, and an objective of this study is to identify feature selection/s method and classifier/s, which are robust and efficient for analysis. The results presented in this section (Fig.2. (a, b)), were obtained after experimenting with the classifiers listed in section 3 and tuning their learning parameters to yield optimum output. The optimal parameters were adjusted by tuning values of batch size, estimator, optimization algorithm, search algorithm, number of iterations, random seed, complexity parameters, weight threshold, etc. The experiments were performed in WEKA 3.8 [11,13]. Firstly, we applied 8 classification algorithms (NB, NN, SVM, RF, LR, K-NN, AdaB, Ensemble) as predictive models, without any feature selection/s, on both Use Case data sets, for determining the functional classes. The accuracy of functional classification covered range from 25% to77% over our Use Cases. The four dominant classifiers providing overall good accuracy were: SVM, LR, NN and RF.

The accuracy of 77.5 %, achieved by SVM at Phylum level of Use Case 1 and accuracy of 50 % by SVM at Species level of Use Case 2; proved to be best prediction results without feature selections. These results proved useful for further comparative analysis. We thereafter, applied both filter based (CFS, IFS, PCA, RB) and wrapper based (LR, SVM, NN, RF) feature selections, using BFS and RM search methods. Overall the combination of Wrapper based filter method with Logistic learner for feature selections' and the classification with LR model {parameter settings: - batch size as 100, with ridge estimator for log likelihood as 1.0E-8}, provided highest accuracy in predicting functional classes for metagenomic studies in hand. The highest accuracy of 97.5 % was attained for MetaPlat rumen data and accuracy of 94.4 % for human microbiota, with the above said combination (Fig.2. (a, b)). The proposed combination, achieved the test average AUC of 0.972 with only 12 OTUs, in comparison to LR, which has AUC of 0.577 with all OTUs (386) in MetaPlat cattle rumen data at genus level of study. Additionally, on human microbiota data, it achieved average AUC of 1.000 with only 4 OTU features, serving much better than LR model having AUC 0.530 over all 756 OTU species.

The application of LR model substantially depicted higher predictive accuracy in comparison to other state of art conventional ML approaches over feature selections. The OTU abundance count data usually have high variance and is not normally distributed. LR proved to be more robust in classification of metagenomic use cases, as it assumes that, the independent OTU features need not to be normally distributed, or have equal variance in each class/functional group.

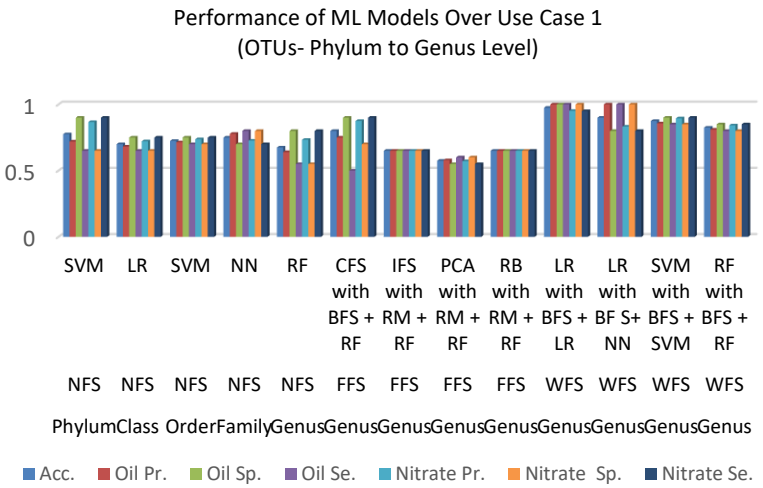

Fig. 2. a. Performance of Classifier/s and Feature Selection/s over Use Case 1

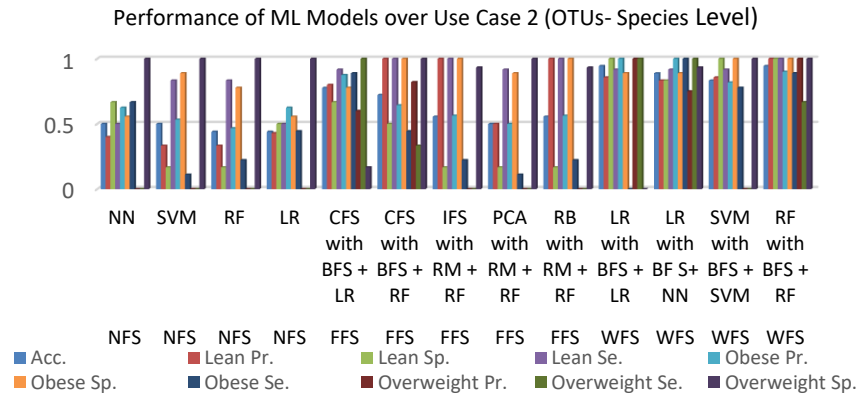

Fig. 3. b. Performance of Classifier/s and Feature Selection/s over Use Case 2 (Here, NFS: No Feature Selection, FFS: Filter Feature Selection and WFS: Wrapper Feature Selection)

The findings report that feature subset selection provides a drive for comparative very good classification accuracy. CFS and Wrapper with LR and RF, proved to be most effective feature selection methods over metagenomic Use Cases. The bacterial sequences were majorly dominated by *Bacteroidetes*, *Firmicutes*, *Actinobacteria*, *Chloroflexi and Proteobacteria* in all ruminants. The significant biological OTU genera features selected by these methods in Use Case 1 are Trichococcus, Tepidimicrobium, Brevibacterium, Methanosphaera, Butyrivibrio, Erwinia and Salana. Bifidobacteriumdentium, Campylobacterconcisus, Helicobacterhepaticus, Mycobacteriummarinum and Borreliaburgdorferi species proved to be significant in analysis over Use Case 2.

5 Conclusion and Future work

In this paper, we have presented an integrative approach to characterize OTU features that are useful in identifying functional roles in metagenomic studies. The results, show that feature selections play an important role in metagenomic analysis. We propose that LR with wrapper based on LR learner considering ridge estimation, potentially give higher validation accuracy for identifying functional roles from human and cattle microbiomes. However, it may be computationally intensive for very large data sets. Also, we considered independent OTU features in LR for metagenomic analysis. In future, we propose to apply new optimization methods with phylogeny-driven LR, integrating association analysis into our framework; for gaining over computational efficiency.

Acknowledgement

This work was supported in part by the MetaPlat project, (www.metaplat.eu) funded by H2020-MSCA-RISE-2015 and Research Strategy Fund of Ulster University.

References

1. Hugenholtz, P. and Tyson, G. W. Metagenomics. *Nature* **455,** 481–483 (2008)
2. McDonald, D. Amanda, B., and Knight, R. "Context and the human microbiome." *Microbiome* 3, no. 1 (2015)
3. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nat. Methods* **5,** 16–18 (2008)
4. Turnbaugh, P. J. et al. A core gut microbiome in obese and lean twins. Nature 457 (2009)
5. Belanche, A. et.al. An integrated multi-omics approach reveals the effects of supplementing grass or grass hay with vitamin E on the rumen microbiome and its function. Front. Microbiol. 7, 1–17 (2016)
6. Roehe, Rainer, et al. Bovine host genetic variation influences rumen microbial methane production with best selection criterion for low methane emitting and efficiently feed converting hosts based on metagenomic gene abundance. PLoS Genet 12, no. 2 (2016)
7. Thomas, T. Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. Microb. Inform. Exp. 2, 3. (2012)
8. Prakash, T., and Taylor. D. Functional assignment of metagenomic data: Challenges and applications, Brief. Bioinform., vol. 13, no. 6. pp. 711–727. (2011)
9. Jonsson, V. Tobias O. et.al. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC genomics 17, no. 1 (2016)
10. Gonzalez, Antonio, and Rob Knight. Advancing analytical algorithms and pipelines for billions of microbial sequences. *Current opinion in biotechnology* 23, no. 1 (2012)
11. Mark, H. Correlation-based Feature Selection for Machine Learning. Methodology. (1999)
12. Kotsiantis, S. B., Zaharakis, I. D. & Pintelas, P. E. Machine learning: A review of classification and combining techniques. Artif. Intell. Rev. 26, 159–190 (2006)
13. Mark, H., Frank, E., Holmes, G. et al., Witten. "The WEKA data mining software: an update." *ACM SIGKDD explorations newsletter* 11, no. 1 (2009)
14. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437 (2009)