Learning from City-specific Metagenomic Fingerprints

Chandrima Bhattacharya[1,*], Debneel Bagchi[1], Lu Wang[2], Somsubhro Mukherjee[2] and Malay Bhattacharyya[1]
[1]Indian Institute of Engineering Science and Technology, Shibpur
Howrah - 711103, West Bengal, India
E-mail: {chandrima.004, debbags}@gmail.com, malaybhattacharyya@it.iiests.ac.in
[2]Department of Biological Sciences, National University of Singapore, Singapore - 119077
E-mail: {Lu.Wang, smukherjee}@u.nus.edu
*** The presenting author**

Abstract

Recent large-scale collection city-specific whole genome shotgun (WGS) metagenomics data has enabled the mapping of microbes available in public spaces and transit systems uniquely. With an explicit focus on metagenomics data from the MetaSUB International Consortium, we study a variety of metagenomic features (including species distribution, relative abundance, diversity, etc.) for the purpose of learning the city-specific fingerprints of different samples. The features are mainly collected using tools available in the COSMOSID platform. With a framework of supervised learning, we highlight that the microbiome community can be linked to its city of origin, in a majority of cases, with a cross-validation accuracy of more than 80%. However for some of the cities (e.g., Ofa and Porto for the MetaSUB data), the principle components are found to be overlapping with other cities, making the classification task difficult with the chosen features. This might be highlighting the necessity of more sophisticated sample collection or feature extraction.

1 INTRODUCTION

Microbiome analysis has recently drawn attention of the worldwide researchers due to its broader scope of understanding higher-level organisms [1]. The microbiome refers to the microorganisms dwelling in a particular environment. Studying the microbiome data, which is increasingly becoming open for analysis, provides us a better picture of the system level activities surrounding an organism. However, it is also interesting to understand the interaction and activities within this microbiome community. The study of diversity in microbiome across the different localities has rarely been studied until recently. The current paper aims to address this issue with the support from MetaSUB International Consortium data. Given the diverse landscape of studies initiated with microbiome data, there is still a scarcity of material-specific analysis of microbiome ecology based on the environments they belong to. Our principal motivation in this paper is to analyze the diversity of microbiome habitat on different materials. We plan to understand the miscellany of different microorganisms in different materials and collected from different sources. The MetaSUB datasets have recently been employed for the purpose of characterizing microbiome community available in the locations like subways [2, 3]. To the best of our knowledge, the study on exploring the city-specific metagenomic fingerprints, which we attempt in the current paper, is in fact novel.

2 DATASET DETAILS

We have basically taken the data of MetaSUB International Consortium for the current analysis. The MetaSUB International Consortium aims to create the world's only longitudinal metagenomic map of mass-transit systems and other public spaces across the globe [4]. The current release of the multi-city analysis data from MetaSUB consists of the microbiome details collected from the cities like Auckland, Hamilton, New York City (NYC), Ofa, Porto, Sacramento, Santiago and Tokyo. A statistical overview of this subset is shown in Table 1. The pilot files are pilot to the study, while the rest of the files were collected on the Global Sampling Day of 21st June, 2016.

Table 1. Details about the subset of MetaSUB training dataset that we analyzed.

	Auckland	**Hamilton**	**NYC**	**Ofa**	**Porto**	**Sacramento**	**Santiago**	**Tokyo**
Continent	Australia and Oceania	Australia and Oceania	North America	Africa	Europe	North America	South America	Asia
Number of samples (Global Sampling Day)	15	16	26	20	60	16	20	20
Number of pilot files	-	-	100	-	-	18	-	-

Our test set consisted of three type of samples: C1, C2Ci and C3 where i = 1, 2 and 3. C1 consisted of 20 mystery samples from the above mentioned 8 cities. The C2Ci included samples from cities from all over the world, outside the above 8 cities. The C3, on the other hand, represent cities which can be from any of the above cities or from a new location not present before. The shotgun data is provided with FASTQ sequences of the microbiomes which are compressed by DSRC(DNA Sequence Reads Compression format) or gunzip. The compressed data size ranges between a few hundred KBs to more than 5GB. It is interesting to note the data has class imbalance with a high bias toward the pilot samples of NYC.

3 METHODS

COSMOSID is a robust tool for quantification and detection of microorganisms [6]. Alongside generating report about classification, it can also produce 3D PCA view, heatmap, matrix, etc. Moreover, with CosmosID, we also get to calculate alpha and beta diversity of the species. CosmosID is widely used by scientists for disease discovery and pharmaceutical research as the curated genome database used is said to provide more than 95% sensitivity and 99% specificity. Other than rapid screening and visualization of microbiomes, CosmosID also provide batch processing and AMR profiles. We use the comparative analysis tool of CosmosID to study the microbiome community for each sample. Features like frequency and relative abundance of bacteria, fungi, protists, virus, respiratory virus, virulence factors, antibiotic resistance, and species diversity are considered in the analysis.

4 RESULTS

We present the details about the empirical analysis in this section. In the beginning, we train the classification model and then test on independent city and continent-specific samples.

4.1 Preliminary Insights

Initially, we classified the samples according to the NCBI taxonomy. We used various visualization present along with the diversity values as features. We have used alpha diversity and beta diversity of the various samples considering domains at particular instance, and for that all analysis have been done by using frequencies of each species. The CHA01, Simpson and Shannon index has been used for calculating α-diversity, while Jaccard and Bray-Curtis for β-diversity.

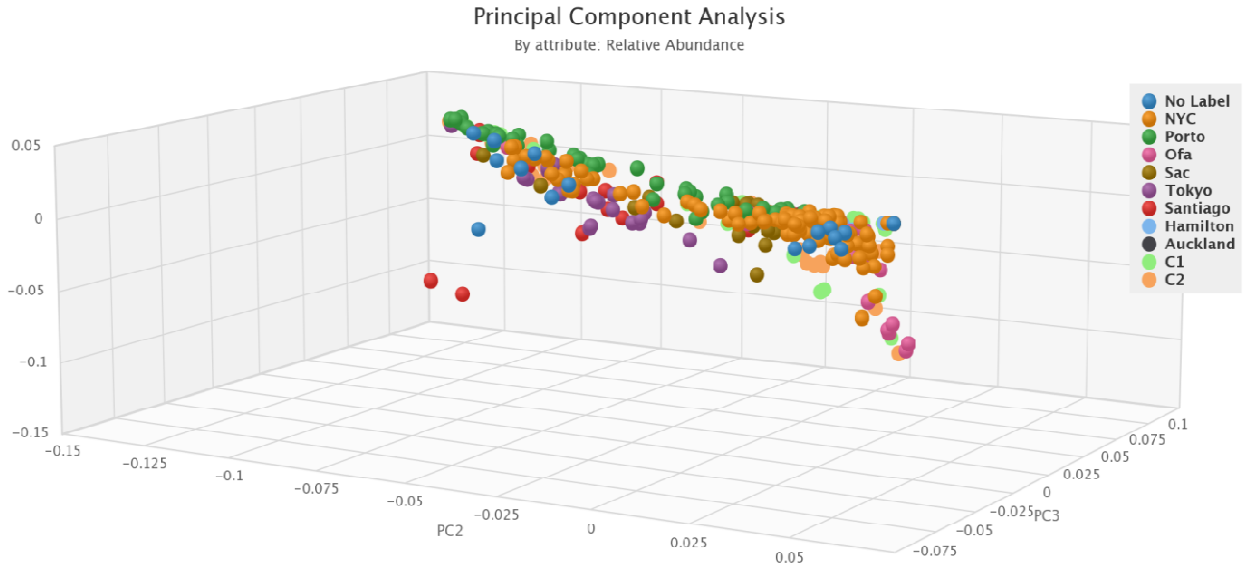

Figure 1: A 3D PCA with respect to its relative abundance of bacteria in each city of origin.

4.2 Learning

The WEKA tool by the University of Waikato, New Zealand was used for classification and prediction of the unknown samples [7]. We chose the matrix table feature containing all the species present along with the number of times each species present (number of hits) to find the local species fingerprint which we used for analysis as dataset. We used a supervised learning algorithm in WEKA, AddClassification to train a learning tree classifier, RandomForest on the whole dataset resulting in which we added two new attributes, classification and error flag to the dataset. The 10 fold cross-validation accuracy is found to be more than 80%. The unsupervised algorithm RemovePercentage was then used to create our training and test datasets with the test set containing all the unknown samples. The test set was then run for classification and we visualized the classifier errors in a new dataset which contained further two new attributes namely, predicted classification and the prediction margin. The error flag provided information regarding the success of our prediction. A higher prediction margin signifies a higher probability for the correct classification (1 being the highest, indicating correct classification with 100% confidence and -1 the lowest, indicating incorrect classification with 100% confidence).

4.3 City Profiling

Using the above learning algorithms on species dataset, we have come to the following predictions for C1 and C3 (see Appendix A and B). We can see that most of the cities are predicted as New York (mostly pilot samples) and Porto. Most of the cases had positive prediction margin. For analysis, we have not calculated the error caused due to the bias in file quantities. The most commonly misclassified files includes New York (mostly pilot samples) as Porto and vice versa. There are few discrepancy between Ofa and Santiago files too with New York files.

4.4 Discrepancy in City Fingerprint

New York had pilot samples from 2015, and samples from 2016 global sampling day. Studying visualization via PCA, we find that both form a distinctive cluster, and that can be even verified using beta diversity PCoA representation or by machine learning. As similar material of collection of data, we find a distinctive variation with respect to time (or maybe season). Hence, a better understanding of temporal as well as seasonal variation is required to understand microbial fingerprints in a more quantitative manner.

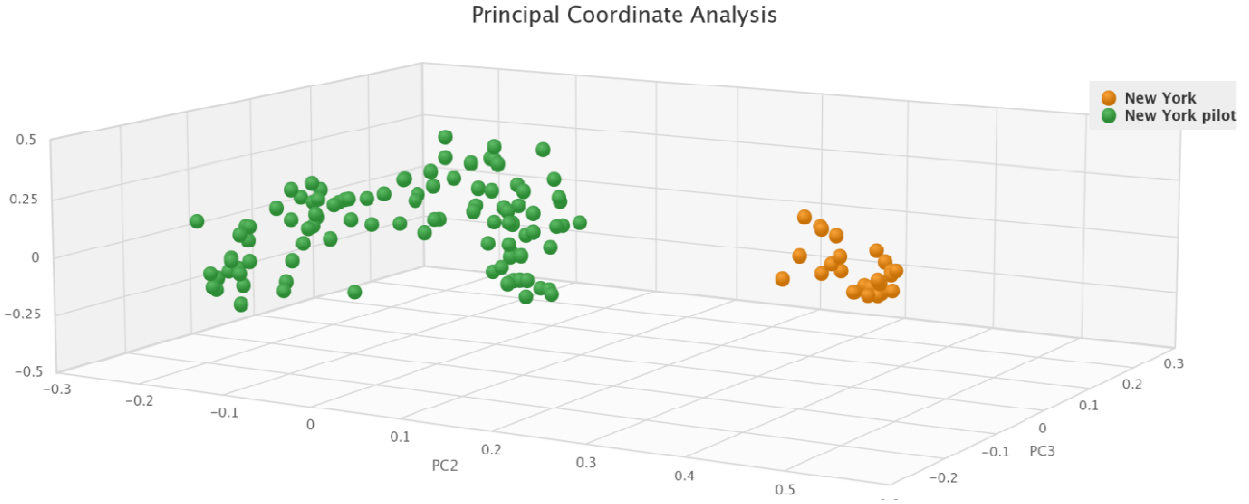

Fig 2: Bacteria 3D β-diversity PCoA for New York samples.

4.5 Continent Prediction

We find that there is no unique microbial fingerprint left in any continent, and there is mostly an overlap. Moreover, the uneven sample distribution is introducing a bias, as there are more files in North America when compared to any other continent. Hence, due to the following reasons, accurate continent prediction is not a viable option. Hence, we have got all the C2 files, C2_C1, C2_C2 and C2_C3, being predicted as North America.

5 CONCLUSION

The current paper provides some interesting highlights about how global mapping can be done based on the capacity of metagenomic fingerprints. A simple prediction model can assist in recognizing the city-specific (and often continent-specific) microbiome patterns. Moreover, we find that studying various microbes (e.g., the effect of respiratory viruses on subway) can be helpful in the future design of smart

cities, by controlling the effects of harmful microbes, thereby preventing major epidemic and endemic breakouts. As a future goal, we realize that diversity parameters (e.g., α-diversity and β-diversity) provide interesting outlooks about microbes. The studies on relation of diversity parameters with respect to city prediction maybe an interesting direction to follow up.

ACKNOWLEDGMENTS

We would like to thank CosmosID Inc. for sponsoring our team. A special thanks to Dr. Nur A. Hasan (CSO,VP CosmosID.) and Manoj Dadlani (CEO, CosmosID) for approving us to use CosmosID for CAMDA 2018 MetaSUB Challenge. Also, we are grateful to Brian Fanelli for helping us with pre-analysis.

REFERENCES

1. I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nature Reviews Genetics*, 13(4):260, 2012.
2. A. D. C. Fernandes, "Characterization of microbiome in Lisbon subway," 2016.
3. X. Triadó-Margarit, M. Veillette, C. Duchaine, M. Talbot, F. Amato, M. C. Minguillón, V. Martins, E. de Miguel, E. O. Casamayor and T. Moreno. "Bioaerosols in the Barcelona subway system," *Indoor Air* , 27(3): 564-575, 2017.
4. MetaSUB International Consortium, "The metagenomics and metadesign of the subways and urban biomes (MetaSUB) international consortium inaugural meeting report," *Microbiome* 4(1):1-14, 2016.
5. Ł. Roguski and D. Sebastian, "DSRC 2 – Industry-oriented compression of FASTQ files," *Bioinformatics* 30.15: 2213-2215, 2014.
6. COSMOSID, Available at: http://www.cosmosid.com
7. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations* 11(1), 2009.

Appendix A: The predicted solution set for C1

1	File Name	Predicted City
2	CAMDA18_MetaSUB_C1_1_R1R2.fastq.gz	Santiago
3	CAMDA18_MetaSUB_C1_2_R1R2.fastq.gz	Santiago
4	CAMDA18_MetaSUB_C1_3_R1R2.fastq.gz	New York pilot
5	CAMDA18_MetaSUB_C1_4_R1R2.fastq.gz	New York pilot
6	CAMDA18_MetaSUB_C1_5_R1R2.fastq.gz	Ofa
7	CAMDA18_MetaSUB_C1_6_R1R2.fastq.gz	New York pilot
8	CAMDA18_MetaSUB_C1_7_R1R2.fastq.gz	Porto
9	CAMDA18_MetaSUB_C1_8_R1R2.fastq.gz	Porto
10	CAMDA18_MetaSUB_C1_9_R1R2.fastq.gz	New York pilot
11	CAMDA18_MetaSUB_C1_10_R1R2.fastq.gz	Porto
12	CAMDA18_MetaSUB_C1_11_R1R2.fastq.gz	Porto
13	CAMDA18_MetaSUB_C1_12_R1R2.fastq.gz	Ofa
14	CAMDA18_MetaSUB_C1_13_R1R2.fastq.gz	Porto
15	CAMDA18_MetaSUB_C1_14_R1R2.fastq.gz	Santiago
16	CAMDA18_MetaSUB_C1_15_R1R2.fastq.gz	New York pilot
17	CAMDA18_MetaSUB_C1_16_R1R2.fastq.gz	New York pilot
18	CAMDA18_MetaSUB_C1_17_R1R2.fastq.gz	Porto
19	CAMDA18_MetaSUB_C1_18_R1R2.fastq.gz	New York pilot
20	CAMDA18_MetaSUB_C1_19_R1R2.fastq.gz	New York pilot
21	CAMDA18_MetaSUB_C1_20_R1R2.fastq.gz	New York pilot
22	CAMDA18_MetaSUB_C1_21_R1R2.fastq.gz	New York pilot
23	CAMDA18_MetaSUB_C1_22_R1R2.fastq.gz	Porto
24	CAMDA18_MetaSUB_C1_23_R1R2.fastq.gz	New York pilot
25	CAMDA18_MetaSUB_C1_24_R1R2.fastq.gz	New York pilot
26	CAMDA18_MetaSUB_C1_25_R1R2.fastq.gz	New York pilot
27	CAMDA18_MetaSUB_C1_26_R1R2.fastq.gz	Porto
28	CAMDA18_MetaSUB_C1_27_R1R2.fastq.gz	Porto
29	CAMDA18_MetaSUB_C1_28_R1R2.fastq.gz	New York pilot
30	CAMDA18_MetaSUB_C1_29_R1R2.fastq.gz	New York pilot
31	CAMDA18_MetaSUB_C1_30_R1R2.fastq.gz	Porto
32		

Appendix B: The predicted solution set for C3

1	File Name	Predicted City
2	CAMDA18_MetaSUB_C3_1_R1R2.fastq.gz	Porto
3	CAMDA18_MetaSUB_C3_2_R1R2.fastq.gz	New York pilot
4	CAMDA18_MetaSUB_C3_3_R1R2.fastq.gz	New York pilot
5	CAMDA18_MetaSUB_C3_4_R1R2.fastq.gz	New York pilot
6	CAMDA18_MetaSUB_C3_5_R1R2.fastq.gz	New York pilot
7	CAMDA18_MetaSUB_C3_6_R1R2.fastq.gz	New York pilot
8	CAMDA18_MetaSUB_C3_7_R1R2.fastq.gz	New York
9	CAMDA18_MetaSUB_C3_8_R1R2.fastq.gz	New York
10	CAMDA18_MetaSUB_C3_9_R1R2.fastq.gz	Sacramento
11	CAMDA18_MetaSUB_C3_10_R1R2.fastq.gz	New York
12	CAMDA18_MetaSUB_C3_11_R1R2.fastq.gz	New York
13	CAMDA18_MetaSUB_C3_12_R1R2.fastq.gz	New York pilot
14	CAMDA18_MetaSUB_C3_13_R1R2.fastq.gz	New York
15	CAMDA18_MetaSUB_C3_14_R1R2.fastq.gz	New York pilot
16	CAMDA18_MetaSUB_C3_15_R1R2.fastq.gz	New York
17	CAMDA18_MetaSUB_C3_16_R1R2.fastq.gz	New York
18		