**2022 Winter Cambridge AI+ Programme**

**Group Project Report**

|  |  |  |  |
| --- | --- | --- | --- |
| **Programme Pathway** | Biotechnology Engineering & Healthcare Technology | | |
| **Course Group** | 22CamAI+ BIOB-G1 | **Group Name** | 42 |
| **Group Leader** | Yixin Guo | | |
| **Group Members** | Ruihan Hou, Bohao Peng, Haohong Qiu, Yufeng Wen | | |

**Predicting Soil Microbial Diversity Based on**

**Soil Properties by AI Approach**

1. **Project Overview**

In the Yellow River Basin area of China, perennial water and soil erosion have led to dramatic changes in the microbial community in this region, and the microbes inhabited in the soil are considered the most abundant and active organisms on the land surface. Prior to our project, soil samples were collected from different sections, and the microbes in each soil sample were characterized by 16s rRNA (for bacteria) and ITS (for fungus) high-throughput sequencing. Based on their data, we managed to figure out the interrelationship between soil properties and microbiomes by some statistical approaches. Furthermore, we constructed an artificial neural network for prediction, in order to better validate this interrelationship.

1. **Project Background**

Globally, soil microbes are the drivers of key biogeochemical cycles involving carbon, nitrogen, phosphorus, iron etc. (Tecon et al. 2017). They are interrelated and constitute the soil microbial community orderly as a whole.

Because different regions or even different parts of the same region can have different physiochemical properties, it can be concluded that geographical differences may lead to different microbial populations. Our limited knowledge of biodiversity and its association with specific physiochemical properties has motivated several soil metagenomics initiatives in the past, the Earth Microbiome Project during 2010-2014 (Gilbert et al. [2014](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5812502/#bib95)) and in 2014 the Brazilian and Chinese Microbiome Projects. Further researches introduced certain tools like network analysis (Banerjee et al. [201](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5812502/#bib184)8), structural equation model (Chen et al. 2013), time series analysis (Karoline et al. 2015) etc., which enabled us to understand the mechanism of this link through traditional statistical approaches. Since the field of microbes have long faced the problem of transferring the achievements from lab-​scale studies to the clinic, field or natural environment (Knight et al. 2018 Márcio et al. 2020), increased need of methods updated with AI development should be taken into consideration. We noticed that there have been researches who focused on the possible applications of machine learning on microbes (Yazdani et al. 2016, Subramanian et.al 2014, Knights et al. 2011), while neural network, a powerful tool in prediction, has not yet arisen much attention in the same field. Only a few researches using neural network to predict microbial communities in acid mine drainage (Kuang et al. 2016) or evaluating benefit of conservation tillage (Hao et al. 2021), but they showed little interest in predicting soil microbial community in erosion area. In this project, we aim to apply AI-based approaches like neural network and hope to gain insight on how neural network performs in this field.

This study was conducted at the Changwu State Key Agro-Ecological Experimental Station in Changwu, Shaanxi, China. The soil was collected from loess deposits and was described as a loam (Cumulic Haplustoll; USDA Soil Taxonomy System) with a clay content of 22% (Wang et al. 2021). Soil samples were randomly collected from the gully head, upstream, midstream, downstream and dam area respectively in Wangdonggou. In total, 25 samples (5 samples from each of the 5 sampling sites) were collected. For each soil sample, six soil cores were randomly collected using a soil auger (d = 3 cm) and then combined into a mixed sample. The samples were placed in a portable refrigerator for transportation to the laboratory. Then, the subsamples were: (1) stored at - 80 ◦C for analyzing the bacteria and fungi composition, and (2) air dried for determining the SOC, TN, Olsen-P.

1. **Division of roles & responsibilities**

**Yixin Guo**

1. Introduced the background and goal of the project, and guaranteed the availability and reliability of the dataset (used for published articles).

2. Read related articles in this area and conjectured the general structure of the project. Discussed about which methods to use and why, suggested where to find useful information about the methods. Made sure that all teammates are aware of their duty.

3. Assigned tasks according to the talent and interest of team members, and organized group meetings regularly.

4. Had discussions with Yufeng Wen in the construction and preparation of neural network.

5. Had discussions with Ruihan Hou about related articles about network analysis. Completed the plotting of network graphs and CCA analysis.

**Yufeng Wen**

1. Read the original article on which our project is based and assisted to introduce the background and techniques used in our project.

2. Learned the basics and general process of neural network construction from many literatures, and learned more detailed information through online tutorials, blog, videos, and through consulting my seniors. Eventually, completed the NN construction of our project.

3. Considered the application of our project and evaluated the results of our project.

4. Revised our PPT and other Word submissions (i.g. Project Proposal & Project Report). Checked the linguistic and logic coherence of our presentations.

5. Tried my best to offer support and assistance technically and spiritually to other team members throughout the project.

**Haohong Qiu**

1. Made Gantt charts, flow charts and completed project Plog and other project tasks according to the team schedule.

2. Contributed to the visualization of the pre-processed data, including the PCoA and UPGMA tree analyses.

3. Learned how to do PCoA and UPGMA tree analysis.

**Peng Bohao**

1. Read the articles about microbial abundance and Alpha diversity. Selected the indices that were necessary to our project.

2. Used GraphPad Prism to visualize the microbial species abundance and microbial alpha diversity.

3. Used SPSS 20.0 for differential analysis of microbial alpha diversity data.

4. Had discussions with team members and completed further analysis of data.

**Hou Ruihan**

1. Read literature about soil microbes. Tried to find a way to show the inherent mechanism of the relationship between the soil properties and the microbes.

2. Learned to perform the Network Analysis (NA).

3. Used R language to do some basic data processing and visualizing.

4. Collaborated with the team and analyzed the visualization results.

1. **Challenges & Solutions**

**1. Describe the diversity of microbiomes**

Alpha diversity refers to the diversity within a particular region or ecosystem and is a composite indicator of richness and evenness. Alpha diversity mainly depends on two factors: species richness and diversity. And we chose Shannon and Simpson to evaluate the microbial diversity in our project. These two indices are proven to reflect species diversity and are often used by researchers to reflect species diversity in studies. Shannon: the richness and evenness of the community were both considered. The higher Shannon index is, the higher community diversity is. Simpson: one of the microbial diversity indices used to estimate samples. It was proposed by Edward Hugh Simpson (1949) and is used in ecology to quantify the biodiversity of certain area. The larger the Simpson index is, the lower community diversity is.

In our data analysis, Shannon index indicated good species diversity, while Simpson index showed poor species diversity. Through literature review, we found that this was because the relative abundance of dominant species in the sample was too high, which led to the increased probability of extraction of dominant species in the calculation of Simpson index.

In beta diversity, PCoA and UPGMA trees are both based on the Bray-Curtis distance to show the relationship between microbial distribution and the environment. In the original dataset, if we didn’t reduce the data dimension, it would be difficult to understand the intrinsic relationship hidden in the data. PCoA and UPGMA tree can both reveal the ecological relationship, and PCoA can be used to further show the clustering among each sample, reflecting a certain ecological gradient.

**2. Reveal the inherent mechanism of changes in microbes**

According to Fath et al., we know that Network Analysis (NA) is a relatively traditional and conventional method to reveal the inherent mechanism of the relation between soil properties and soil microbes. Its applications in ecology field are called Ecological network analysis (ENA), which is a systems-oriented methodology to analyze within system interactions used to identify holistic properties that are otherwise not evident from the direct observations. So, we want to apply this method to our data. We use R to process the data and visual the network. From the figures we know the bigger the node is, the more interaction it has with other microbes. So, we can easily tell the biggest node which is called keystone taxa. We can also easily know the correlations from the clarity of the links, which is called community structure. Besides from network analysis, we will further use CCA to reveal the mechanism of how related properties influence microbial community.

**3. Construct the artificial neural network**

Due to our major barrier, the construction of our artificial neural network (ANN) had to begin from the basics, thus posing a lot of challenges to us. To tackle this problem, we turned to online tutorials[[1]](#footnote-1) , and videos[[2]](#footnote-2) for help and consequently had a better understanding of ANN construction by PyTorch.

We decided to use Python for our ANN construction, since it was said to be a pervasive high-level programming language (Rahman et al. 2021), and can support a wide range of frameworks (Johansson et al. 2012). And we initially chose PyTorch as our framework, due to its powerful calculation ability described in the work of Waqed et al. Moreover, because of the simplicity and dearth of our data, we decided to construct a simple ANN model with only one hidden layer. Then, with the sketch of our ANN well-depicted, we moved on to further implement this prediction model. During this process, we’ve firstly been troubled by the selection of activation function. In online videos by Morvan\_Zhou[[3]](#footnote-3), it was said that for the simplest model with only one hidden layer, nearly all activation functions (i.g. ReLu, Sigmoid, Tanh, etc.) could be applied to fit in. But later, by discussing with our senior Dr. Zhengwei Tao, we finally decided to use Sigmoid as the activation function of our ANN model since our output (microbial relative abundance) is ranging between 0 and 1, which is precisely in accordance to the output range of Sigmoid function. Other challenges we’ve met in constructing ANN models were relatively easier and were all solved by us using similar methods as described above.

1. **Contributions & Limitations**

**A) Contribution**

1. Many research before focus on getting to know statistical correlation between physiochemical properties and soil microbial community and conclude what may be the main factor influencing the latter. Besides concluding the crucial factor, our research seeks for predicting relative abundance of keystone taxa in microbial community in Wangdonggou.

2. Two approaches, including the traditional statistical approach and the novel ANN approach, were both implemented in our project to analyze the interrelationship between soil and soil microbes. The mean square error of our ANN model showed that it had performed well in predicting soil microbes given certain soil properties.

**B) Limitation**

1. Because of the compositional nature, we cannot change the proportion of one bacterial species without changing that of at least one other because the proportions must sum to 1. That property renders many standard multivariate statistical methods inappropriate or inapplicable. For example, with only direct calculation of Pearson’s correlation between two components of compositional data can lead to spurious correlations.

2. In addition, bias can arise from the cumulative effect of both systematic and random errors throughout the whole process, including DNA extraction, sampling, amplicon, sequencing, and bioinformatics. Moreover, each step of a metagenomics workflow (DNA extraction, PCR, and sequencing) adds an error that seems to be taxa-specific. Friedman and Alm (2012) observed that when the population contains dominant taxa, the problem of spurious associations is very severe.

3. Since this high-throughput sequencing technology is still not that cheap, our dataset remains relatively small in sample size. In this project, we do not broaden our research dataset due to the time limitation, thus, the application of neural network may not be that appropriate.

**Bibliography**

Wang, R, Wang, Y, Zheng, W, Hou, F, Hu, Y, Guo, S (2021). Converting croplands to orchards changes soil microbial community composition and co-occurrence patterns. *Land Degrad* 2509– 2519.

Tecon, Robin, and Dani Or (2017). “Biophysical processes supporting the diversity of microbial life in soil.” *FEMS microbiology reviews* vol. 41,5: 599-623.

Banerjee, S., Schlaeppi, K, M. G. (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology*, 16(9), 567-576.

Chen, D., Lan, Z., Bai, X., Grace, J.B., Bai, Y., (2013). Evidence that acidification-induced declines in plant diversity and productivity are mediated by changes in below-ground communities and soil properties in a semi-arid steppe grassland. *J. Ecol.*101, 1322–1334

Wang, G., Koziol, L., Foster, B. L., & Bever, J. D. (2022). Microbial mediators of plant community response to long-term N and P fertilization: Evidence of a role of plant responsiveness to mycorrhizal fungi. *Global Change Biology*, 00, 1–15.

Gilbert JA, Jansson JK, Knight R (2014). The Earth Microbiome project: successes and aspirations. *BMC Biol*; 12:69. Published 2014 Aug 22.

Knight, R., Vrbanac, A., Taylor, B.C. et al. (2018) Best practices for analysing microbiomes. *Nat Rev Microbiol* 16, 410–422.

Karoline Faust, Leo Lahti, Didier Gonze, Willem M de Vos, Jeroen Raes. (2015) Metagenomics meets time series analysis: unraveling microbial community dynamics, *Current Opinion in Microbiology*, Volume 25, Pages 56-66, ISSN 1369-5274

Márcio F.A. Leite, Eiko E. Kuramae. (2020). You must choose, but choose wisely: Model-based approaches for microbial community analysis, *Soil Biology and Biochemistry*, Volume 151, 108042, ISSN 0038-0717,

M. Yazdani, B. C. Taylor, J. W. Debelius, W. Li, R. Knight and L. Smarr. (2016).Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease *2016 IEEE International Conference on Big Data (Big Data)*, pp. 1272-1280

Subramanian, S., Huq, S., Yatsunenko, T. et al. (2014) Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* 510, 417–421.

Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat Methods*. 761-3.

Kuang, J., Huang, L., He, Z. et al. (2016) Predicting taxonomic and functional structure of microbial communities in acid mine drainage. *ISME* J 10, 1527–1539.

Jiaqi Hao, Yue Lin, Guangxin Ren, Gaihe Yang, Xinhui Han, Xiaojiao Wang, Chengjie Ren, Yongzhong. Feng(2021),Comprehensive benefit evaluation of conservation tillage based on BP neural network in the Loess Plateau, *Soil and Tillage Research*, Volume 205.

Fath, B. D., Scharler, U. M., Ulanowicz, R. E., & Hannon, B. (2007). Ecological network analysis: network construction. *Ecological modelling*, 208(1), 49-55.

Rahman T.H., I. S. (2021). A Timeline from Neural Network to Python Through ANN and AI: An Introductory Tutorial. *Neuro-Systemic Applications in Learning*.

Johansson J.R., N. P. D., Franco N. (2012). QuTiP: An open-source Python framework for the dynamics of open quantum systems. *Computer Physics Communications,* 183**,** 1760-1772.

Waqed H. Hassan, H. H. H., Musa H. Alshammari, Halah K. Jalal, Sajjad E. Rasheed (2022). Evaluation of gene expression programming and artificial neural networks in PyTorch for the prediction of local scour depth around a bridge pier. *Results in Engineering,* 13.

Friedman, J., Alm, E.J. (2012). Inferring correlation networks from genomic survey data. *PLoS Computational Biology,* 8, 1002687.

**FAQ**

**What are the format requirements for the Project Report?**

The format for a project report varies between different disciplines, but all good reports must be both informative and persuasive. Please follow the following general format requirements:

* + Word Count: 2,500 (Max); 2000 (Min)
  + Font Size: 15 (Max); 9 (Min)
  + Line Spacing: 2 (Max); 0.8 (Min)
  + Citation Format: Harvard Citing and Referencing Style ([click here](http://www.imperial.ac.uk/admin-services/library/learning-support/reference-management/harvard-style/)for more information)

**Statement on Plagiarism**

Make sure that you acknowledge the authors of ALL publications you use to write your report. Failure to do so will be considered as plagiarism. Do not copy word for word what an author has said. You may think that the original author has presented the information using the best possible words in the best format. However, it is best to analyse the information presented and re-write it in your own words. If you absolutely have to quote an author ad verbatim, then make sure that you use quotation marks and italics to indicate it.

1. <https://www.dataquest.io/blog/jupyter-notebook-tutorial/> [↑](#footnote-ref-1)
2. https://www.bilibili.com/video/BV1Vx411j7kT?p=1&share\_medium=android&share\_plat=android&share\_session\_id=6dc2c189-c975-431d-b59d-5aacb3d749e9&share\_source=WEIXIN&share\_tag=s\_i&timestamp=1644316517&unique\_k=0LzmR1f [↑](#footnote-ref-2)
3. <https://www.bilibili.com/video/BV1Vx411j7kT?p=10> [↑](#footnote-ref-3)