

The Relative Abundance Analysis of Microbial Community in the Baltic Sea

Tria Puspa Sari¹

Department of Mechanical Engineering, Faculty of
Engineering
University of Pembangunan Nasional “Veteran” Jawa
Timur
Surabaya, Indonesia
tria.puspa.tm@upnjatim.ac.id

Torbjörn E. M. Nordling²

Department of Mechanical Engineering, Faculty of
Engineering
National Cheng Kung University
Tainan, Taiwan (R.O.C)
martin@mail.ncku.edu.tw

Abstract— The microbial community is important to control the sustainability of water ecosystems. The microbial community also has a negative impact on the environment. In particular, the toxin of Algae Bloom in the Baltic Sea which is harmful to the environment and humans. Several previous studies have tried various machine learning methods to analyze algae growth, but with limited achievement. The 167 samples were collected from 2012 to 2013 by the Linnaeus Marine Observatory at 11 km off the shore of Kårehamn in the Baltic Sea. Here we analyze the abundance of the microbial community using the Operational Taxonomic Units (OTUs) based on the normalization method after filtering samples. The aim to analyze the composition of the microbial community based on the past measurement of microbial composition. It is the pre-processing of data which is the step toward the prediction in the future. The result shows large different variations in each of OTUs in the fractions from the same sample. Many OTUs that are very abundant in one fraction but very rare in the other fraction. The Large difference abundance of OTU composition make a major challenge when predicting OTU composition in the future. Further quantification of normalization method is required in the pre-processing of data to get the proper data for prediction.

Keywords—Microbial community; Normalization data; Relative abundance, OTUs

I. INTRODUCTION

The microbial community is group of microorganisms which that live together in the same place [1]. They are play a meaningful role in the water ecosystems. They also support in many productions, such as for producing yogurt, cheese, soybean sauce, biomass, fossil fuel, etc [2]. Baltic Sea is a large brackish sea that has high diverse microbes, hence it will be interesting to study since the unique and diverse [3]. The Baltic Sea is also the second largest of brackish water in the world that has an area of 415 000 km² and 85 million people live in surrounding [4].

Furthermore, microbial community has complex structures and patterns. Therefore, it needed to analyze the structures and patterns of microbial community, especially, the analyze of microbial community using the composition of Operational Taxonomic Units (OTUs). OTUs are showing with the read of OTUs in each sample by numerical values. For example, the

number of 0 (zero) shows the absence of OTUs and the number of more than 0 means the presence of OTUs in the sample, and they are presented in a table [5].

Previously, many researchers have tried spatial changes study in the diversity of microbial community in water ecosystems [6][7][8][9]. The microbial community composition has diverse spatial scale and environments [10]. Hence, it needed to consider in microbial community composition which it could defined to heterogenous and homogenous of microbial community composition. Because of the importance of the microbial community in the water ecosystems, many researchers study this field [11]. However, advanced microbial community analysis is required to get a satisfying result.

The sample of microbial data was taken from the sea and was processed by the biological laboratory process. The first step was filtering samples through the filters with specific scales (Fraction). After that, the water samples were done by another laboratory process such as DNA extraction, library preparation, and sequencing process to get the OTUs composition.

The filtering process could give the change of microbial community effects. This effect has happened in most of the microbial studies. The microbial community diversity could be affected by the filter types and the size of filter particles. So, the number of microbial community changes in the filtering process [12]. Many factors could be affected in the sampling then the composition of the microbial community to be heterogeneous in freshwater ecosystems. The factors are the distribution limitation, patchiness of property, water flow, and interaction of species in the local community which depend on the spatial scale [6].

The common issue of the microbial community is the heterogeneity of the composition. The factors of heterogeneity are patchiness and stable stratification [13]. The depth of collecting the water samples also might affect the heterogeneity. The homogeneity of the composition can happen in horizontal location sampling (at a fixed depth) between different locations. However, the composition of the

community could be heterogeneous in vertical location sampling.

Theoretically, the microbial community with heterogeneous specifics has happened on small-scale and or large-scale filtering. Therefore, we expect that this study has the homogeneous specification of the microbial community because of the medium scale of filtering. Thus, in this study will focus on proving the diversity that this microbial community has in relative abundance analysis.

II. METHODS

A. Description of data

The samples were collected twice weekly in 2012 to 2013 at the Linnaeus Microbial Observatory (LMO), N 56°55.851,E 17°03.640, at 2 m depth, 11 km offshore Kårehamn, Baltic Sea by our collaborators at Science for Life Laboratory (SciLifeLab). The dataset is the number of OTU presence in each sample were taken. The raw dataset consists of 8 094 OTUs and 477 data points that taken at 167 time-points (day samples). There are few days that two or more sample were taken from the same place point.

TABLE I. PARTIAL RAW DATASET MATRIX SHOWING THE PRESENCE OF OTUS IN EACH SAMPLE

OTUs	477 data points				
	P1202_02	P1202_03	P1202_04	...	P1851_518
OTU_55	1	0	1	...	1
OTU_56	1	0	0	...	0
OTU_77	6	2	1	...	26
...	...	...	...	...	...
OUT_45 078	0	0	0	...	0

TABLE I. shows the raw dataset that consist the number of read of the OTU in each sample. The row header represents the OTUs and the column header represents the sample identities. The values represent the reads of OTUs in each sample. For example, OTU_55 only present one time in sample of P1202_02 and OTU_77 present 26 times in sample of P1851_518. Lot of OTUs are absent in every sample so there are many zero in the table.

This study uses three fractions to separate data in the filtering process. Fig 1. is showing the three kinds of filters that are used to isolate the sample based on the size of the filters. The water sample is 20 liters (8 094 OTUs) was flowing to filter with a 30 μ m, then they are separated into two filter sizes, there is 0.2 μ m and 3 μ m. Thus, we named with Fraction 0.2-30 for a filter of 0.2 μ m and Fraction 3-30 for a filter of 3 μ m. Then, the samples which are trough of Fraction 3-30 will flow to another 0.2 μ m filter, Fraction 0.2-3.

This figure illustrates the intersection of OTUs and shows the number of OTUs presented in Fraction 3-30 which is less than other fractions. Because Fraction 3-30 has the biggest size hole of the filter than other fractions. Thus, the number of

OTUs that were trapped in Fraction 3-30 is 7 279 because most of the size of the microbial is less than 3 μ m. Fraction 0.2-30 has 8 093 OTUs that are trapped, it means only one OTU has a size of less than 0.2 μ m. Fraction 0.2-30 has 7 594 OTUs that were trapped in the filter after through Fraction 3-30.

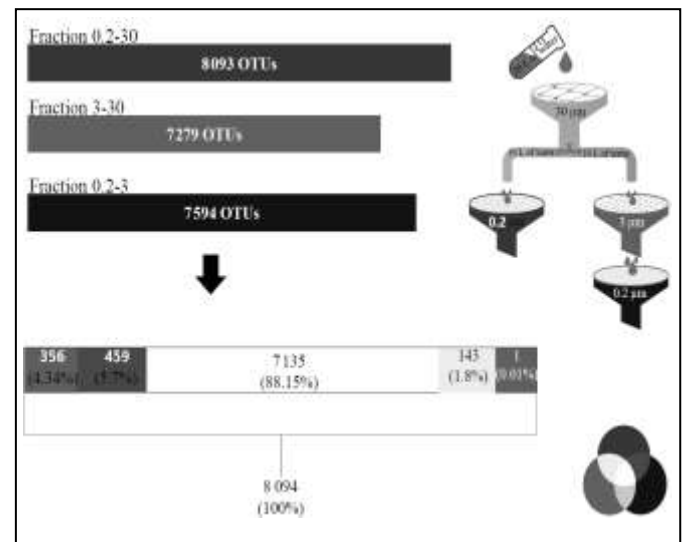


Fig. 1. Illustration of intersection of OTUs between different fractions.

B. Timeline of sampling

The water samples collected over three years during 167 days (data-points) are presented in Fig. 2. This figure illustrates the different numbers of data-points in every year, which is the sample on different days. In 2011, there are 46 samples taken from the location sampling, and there are 61 and 60 samples in 2012 and 2013. The white dot represents the time-points, and the numbers indicate the distance between consecutive samples taken in these years. The length of the line represents the number of days between sampling dates.

C. Normalization Method

The normalization of data is purpose to structuring database in according the series of normal forms in order which are consists of reduce and improve the quantity of data. This is needed to calculate differences of each sample that are present from the measurement of microbial communities [14]. The normalization method could be reducing the biases of the samples, particularly the biases of microbial data that has large biases.

The relative abundance analysis uses one of the normalization methods, it is the total sum normalization (TSS). TSS is the most common method used for normalization. This method is divide each data point by the total of the sample to get the normalized data in the range between 0 and 1. In this study, this method applies to dividing each read of OTU with the total read of OTUs in a sample. Thus, we can see the ratio scale of feature by the total sum of OTUs in the sample. The relative abundance is the proportional abundance of species in the sample. The function of TSS following below [14]:

$$TSS(w^{(j)}) = [w_1^{(j)}/m^j, w_2^{(j)}/m^j, w_3^{(j)}/m^j, \dots, w_p^{(j)}/m^j] \in S^{n \times p} \quad (1)$$

where, TSS ($w^{(j)}$) is a normalization values in each data point divided by total sum of a sample, ($w_p^{(j)}$) is feature or data point, n is sample and p is a the number of data point.

Commonly, the relative abundance is used to shows how common or rare an individual species in the community. Then, to know the relation of species to other species in a sample [15].

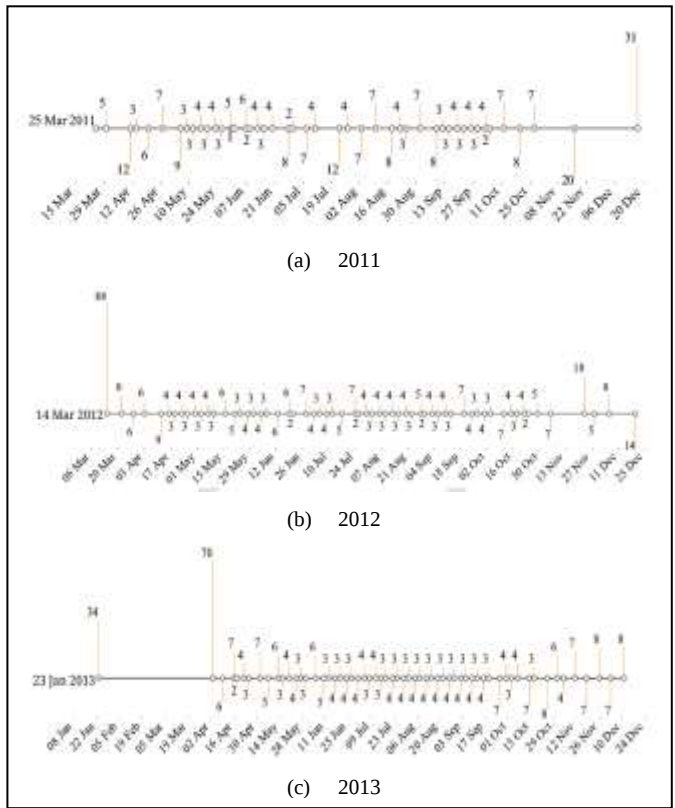


Fig. 2. Distance between consecutive samples in (a) 2011, (b) 2012, and (c) 2013. Each dot represents the dates when the samples were taken and the length of line represents the number of days between consecutive samples.

III. RESULTS AND DISCUSSION

The purpose of relative abundance analysis is to show the relation of OTUs within each sample, also showing the comparison of the microbial community with each other. Thus, in this study, we analyzed the relative abundance and distribution of OTU composition to show the diversity of the microbial community in the Baltic Sea. The standard calculation was applied to know the relative abundance of the microbial community within the use of simple proportions or using rarefying of counts [16]. However, both of these calculations are unsuitable for the disclosure of the differentially abundant community.

The read count of OTUs is a very raw measure since we did not notice the relative amount of microbial caught in each filter, which will also vary during the year, such as providing on the relative abundance of filamentous microbial. On other hand, we would use the option of relative abundance numbers to normalize the data so showing the distribution of OTU composition in the samples. Few studies also use the relative abundance of OTU to present the number of OTUs within detected in the sample in percentage [17].

A. Relative Abundance based on Read Count

The relative abundance is one of normalization method which is obtained by the total sum normalization calculation. Divide the read count of each OTU by the total read count of the OTUs in the sample, then the relative abundance was gotten it. The relative abundance value is in range between zero to one (0-1). This value is almost the same as other normalization calculations.

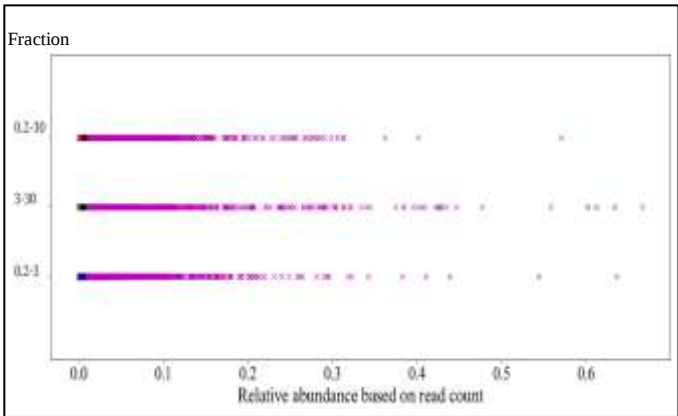


Fig. 3. Scatter plot comparing the relative abundance of the three fractions.

Fig. 3 is presenting the relative abundance values of each OTU in three fractions without care for the number of data points. The goal of this plot is to compare the relative abundance values of each fraction. The relative abundance of each fraction is a range between 0 and 0.7 that the most values close to zero and only a few OTUs close to 0.7. The calculation of relative abundance is divide the each OUT with the total sum of read counts in each data point. The sum of the relative abundance values will equal the number of samples in each fraction. The mean value (square) and median value (circle) of each fraction is marked. The relative abundance of most OTUs is very low. Considerable differences among the OTUs present in high abundance among the fractions exist.

TABLE II. RELATIVE ABUNDANCE STATISTICS

Fraction	Statistic calculates			
	Min	Median	Mean	Max
0.2-30	0	1.9×10^{-3}	6.6×10^{-3}	0.57
3-30	0	1.2×10^{-3}	5.0×10^{-3}	0.67
0.2-3	0	1.3×10^{-3}	5.1×10^{-3}	0.64

TABLE II. refers to the Fig. 3 is showing the relative abundance statistic for each fraction. There is the minimum, median, mean, and maximum of the relative abundance values. The minimum relative abundance in each fraction is zero because there is a lot of OTUs absence (no present) in each sample. In other words, there is a lot of the read of OTUs is 0.

The median and mean of relative abundance are close to zero in each fraction. It has happened because there are many values of relative abundance under 0.1. There are a lot of the read of OTUs shows in small numbers, and only a few in the high number of OTUs read. Thus, the relative abundance maximum is 0.667 in Fraction 3-30, 0.636 in Fraction 02.3-3, and 0.571 in Fraction 0.2-30.

B. Distribution of Relative Abundance of OTUs over Data Points

Fig. 4 describes the relationship between the point of relative abundance and read count of two OTUs in the same sample. Here, they illustrate the relation of the read count of two OTUs in the relative abundance of Fraction 3-30 by comparing the values. We assume Point A means the relative abundance magnitude 10 000 times of the real relative abundance from two different OTUs. It means to make it easy to read and analyze the relation of both of them. At the same time, Point B illustrates the read count of OTUs in two different OTUs. The x-axis represents the OTU_1134, and the y-axis is OTU_23244. These OTUs randomly choose. Hence, the figure is showing the two vectors of the relative abundance and read count overlap each other as they should when scaled.

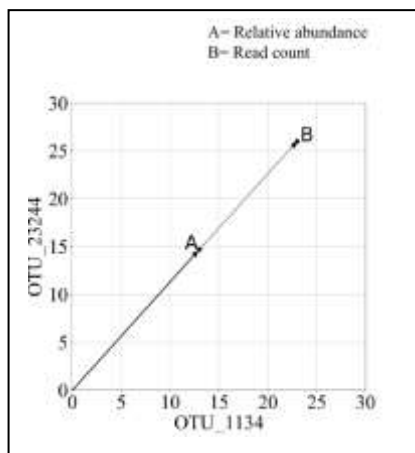


Fig. 4. Illustration of how relative abundance by the total sum of read counts correspond to a pure scaling of the values.

Fig. 5 is showing the distribution of relative abundance in each fraction, (a) Fraction 0.2-30, (b) Fraction 3-30, and (c) Fraction 0.2-3. The figures represent the number of relative abundance across the number of data points each fraction. We can compare the three graphs that there are most of the relative abundance values close to zero. Many OTUs do not exist in the samples and only a few OTUs present in a sample with a high OTUs read. Those happen because of the relative abundance of OTUs affected by the number of reads count of OTUs as significant values.

Hence, there is a gap in the trend of relative abundance in the samples. The distribution shows the importance of the normalization of data to control the few high read of OTUs values with other OTUs. The function of the normalization of data is to make the samples comparable under different assumptions. In these cases that the total number of OTUs reads could be constant in all samples. Thus, the data are proper to be used for predicting the composition of OTUs in the future.

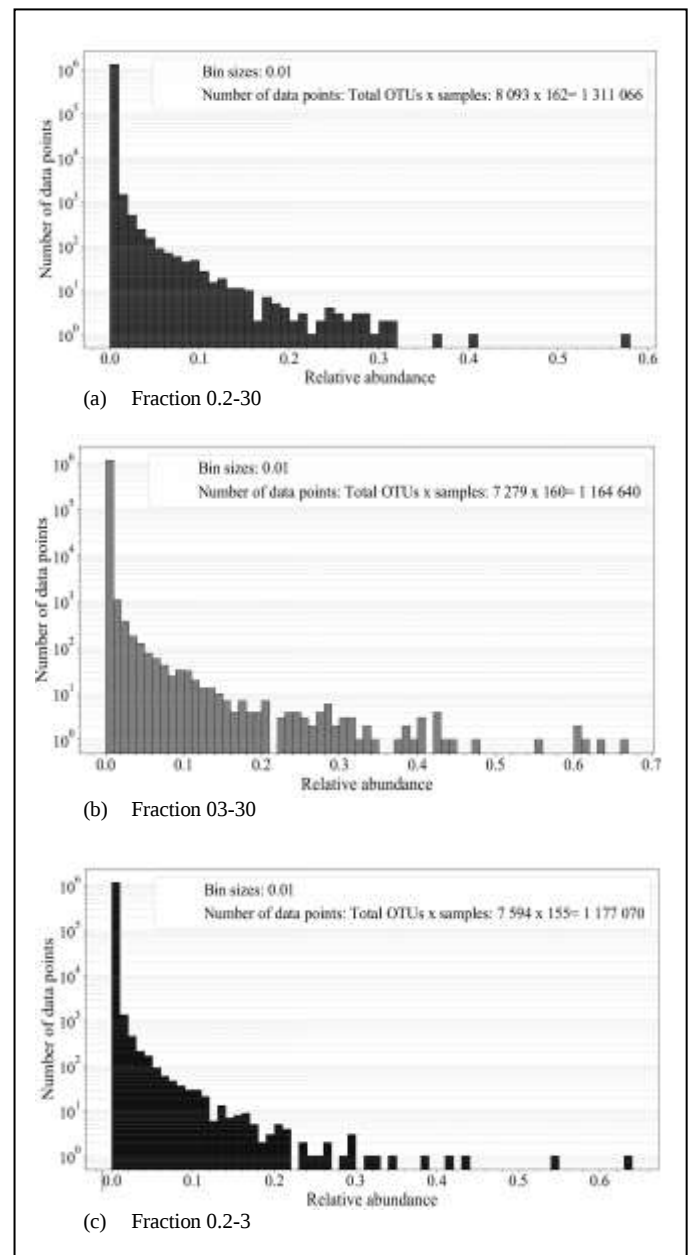


Fig. 5. Histogram of the distribution of relative abundance of OTUs over all data points, i.e. no distinction is made between days or OTUs, for Fraction 0.2-30 (a), Fraction 3-30 (b), and Fraction 0.2-3 (c). The three histograms have similar shape, but Fraction 3-30 contains a few OTUs that for some sample is more abundant than OTUs in the other fractions.

IV. CONCLUSION

The study wanted to quantify the variation of relative abundance to show that this research has differences between fractions. The purpose to prove that Baltic Seawater contains a heterogeneous mixture of the microbial community. Many OTUs are very abundant in one fraction but very rare in the other fraction. Thus, these large differences in the abundance of OTUs make a major challenge when we will predict OTUs in the future based on the composition of OTUs based on past measurements.

In conclusion, when this study wants to predict OTU presence and abundance, this study does not know which of the OTUs should be predicted and which one is showing the exact condition in the sea. We need to quantify with other quantification methods to know the natural variation in the water sea. It means that to quantify to show the natural level of patchiness in the microbial community. The result is showing a significant variation of relative abundance between fractions. Thus, we do not know what is causing that variation. Further, this is a major problem in the prediction because we can not choose which one to predict when the study wants to predict it in the future.

ACKNOWLEDGMENT

The authors would like to acknowledge our collaborators at Science for Life Laboratory (SciLifeLab).

REFERENCES

- [1] A. Konopka, "What is microbial community ecology," *ISME J.*, 2009.
- [2] A. E. Magurran, *Measuring biological diversity. Edition Illustrated.* 2004.
- [3] K. Ininbergs, B. Bergman, J. Larsson, and M. Ekman, "Microbial metagenomics in the Baltic Sea: Recent advancements and prospects for environmental monitoring," *Ambio*, 2015.
- [4] C. Rönnerberg and E. Bonsdorff, "Baltic Sea eutrophication: area-specific ecological consequences," *Hydrobiologia*, vol. 514, no. 1–3, pp. 227–241, Feb. 2004.
- [5] J. O. Corliss, P. H. A. Sneath, and R. R. Sokal, "Numerical Taxonomy: The Principles and Practice of Numerical Classification," *Trans. Am. Microsc. Soc.*, 1974.
- [6] C. A. Hanson, J. A. Fuhrman, M. C. Horner-Devine, and J. B. H. Martiny, "Beyond biogeographic patterns: processes shaping the microbial landscape," *Nat. Rev. Microbiol.*, vol. 10, no. 7, pp. 497–506, Jul. 2012.
- [7] N. Fierer *et al.*, "Cross-biome metagenomic analyses of soil microbial communities and their functional attributes," *Proc. Natl. Acad. Sci. U. S. A.*, 2012.
- [8] S. M. Gibbons, J. G. Caporaso, M. Pirrung, D. Field, R. Knight, and J. A. Gilbert, "Evidence for a persistent microbial seed bank throughout the global ocean," *Proc. Natl. Acad. Sci. U. S. A.*, 2013.
- [9] J. Whitaker *et al.*, "Microbial community composition explains soil respiration responses to changing carbon inputs along an Andes-to-Amazon elevation gradient," *J. Ecol.*, 2014.
- [10] J. L. Liang *et al.*, "Fine-scale spatial patterns in microbial community composition in an acid mine drainage," *FEMS Microbiol. Ecol.*, 2017.
- [11] L. W. Hugerth and A. F. Andersson, "Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing," *Front. Microbiol.*, vol. 8, no. SEP, pp. 1–22, 2017.
- [12] S. Taguchi and E. A. Laws, "On the microparticles which pass through glass fiber filter type GF/F in coastal and open waters," *J. Plankton Res.*, 1988.
- [13] K. Koskinen, J. Hultman, L. Paulin, P. Auvinen, and H. Kankaanpää, "Spatially differing bacterial communities in water columns of the northern Baltic Sea," *FEMS Microbiol. Ecol.*, 2011.
- [14] M. Badri, Z. Kurtz, C. Muller, and R. Bonneau, "Normalization methods for microbial abundance data strongly affect correlation estimates," *bioRxiv*, 2018.
- [15] G. Jetschke and S. P. Hubbell, "The Unified Neutral Theory of Biodiversity and Biogeography," *Ecology*, 2002.
- [16] S. Bessette, Y. Moalic, S. Gautey, F. Lesongeur, A. Godfroy, and L. Toffin, "Relative abundance and diversity of bacterial methanotrophs at the oxic-anoxic interface of the congo deep-sea fan," *Front. Microbiol.*, 2017.
- [17] S. Wu, G. Wang, E. R. Angert, W. Wang, W. Li, and H. Zou, "Composition, diversity, and origin of the bacterial community in grass carp intestine," *PLoS One*, 2012.