Co-occurrence network analysis of biogas reactor microbiomes based on metagenomics sequencing data

Youwei Huan

Independent project • 30 hp
Swedish University of Agricultural Sciences, SLU
Department of Animal breeding and genetics
Animal Science - Master's Programme

Co-occurrence network analysis of biogas reactor microbiomes based on metagenomics sequencing data

Youwei Huan

Supervisor:	Erik Bongcam-Rudloff, SLU, Department of Animal breeding and genetics
Assistant supervisor:	Bettina Müller, SLU, Department of Molecular Sciences
Examiner:	Bengt Guss, SLU, Department of Biomedical Science and Veterinary Public Health

Credits:	30 hp
Level:	Second cycle, A2E
Course title:	Independent project in Animal Science
Course code:	EX0870
Programme/education:	Animal Science - Master's Programme
Course coordinating dept:	Department of Animal breeding and genetics

Keywords: Co-occurrence network, Metagenomics, Anaerobic digestion

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal science
Department of Animal Breeding and Genetics

Publishing and archiving

Approved students' theses at SLU are published electronically. As a student, you have the copyright to your own work and need to approve the electronic publishing. If you check the box for **YES**, the full text (pdf file) and metadata will be visible and searchable online. If you check the box for **NO**, only the metadata and the abstract will be visible and searchable online. Nevertheless, when the document is uploaded it will still be archived as a digital file.

If you are more than one author you all need to agree on a decision. You can find more information about publishing and archiving here: https://www.slu.se/en/subweb/library/publish-and-analyse/register-and-publish/agreement-for-publishing/

☒ YES, I/we hereby give permission to publish the present thesis in accordance with the SLU agreement regarding the transfer of the right to publish a work.

☐ NO, I/we do not give permission to publish the present work. The work will still be archived and its metadata and abstract will be visible and searchable.

Abstract

Metagenome sequencing makes it possible to capture microbial communities in their diversity, to investigate population structures and evolutionary relationships, and to predict metabolic functions and interactions within the microbial community and with the environment. Metagenomics thus, significantly expands the research scope of microbiology. Compared with 16S sequencing, metagenome sequencing can get more information of species identification, and in-depth research on gene and function. Metagenome-assembled genomes (MAGs) retrieved from metagenomics data of 20 AD biogas reactors and the respective Illumina sequencing reads were used to create an abundance table and conduct co-occurrence network analysis. A feasible method was designed for making OTU table and network analysis of metagenomics data and some potential clusters were found may connected to the food chain in microbial community.

Keywords: Co-occurrence network, Metagenomics, Anaerobic digestion

Table of contents

List of figures

Abbreviations

AD	Anaerobic Digestion
OTU	Operational Taxonomic Units
MAGs	Metagenome-assembled genomes
TPM	Transcripts Per Million
ASV	Amplicon Sequence Variants
PTS	phosphotransferase systems
CO_2	Carbon Dioxide
DNA	Deoxyribonucleic acid
H_2	Hydrogen

1. Introduction

1.1. Metagenomics

Microorganisms are a vital part of ecosystem, and efforts to study the deeper mechanisms of microorganisms through laboratory culture have never stopped, but the achievements are limited (Hugenholtz *et al*. 1998). The emergence of metagenomics changed the situation. Metagenomics studies avoids the limitations of microbial isolation and pure cultures, enables the discovery of microbial resources, and provides an effective tool for studying microbial communities (Riesenfeld, Schloss, and Handelsman 2004). Metagenomics sequencing gives insights in the diversity of microbial community, population structure, evolutionary relationships, metabolic function, collaborative microbial interactions and its relationship to the environment. This greatly expands the research scope of microbiology (Handelsman 2004). Metagenomics gives great opportunities to gain insight into the biology of those microorganisms that cannot be cultured in the laboratory and therefore remain unknown. The detailed understanding of these unknown microorganisms can help restore authentic microbial communities in various situations (Daniel 2005). Compared to 16S rRNA sequencing, metagenomics whole genome sequencing can get more information of species identification and enables to carry out in-depth research on genes and functions. More detailed species identification information can also give better understanding of microbial communities.

1.2. Anaerobic digestion

Anaerobic digestion (AD) has been widely used in converting sewage, livestock and poultry manure, and urban organic waste into methane and some other fuels and fertilizers. Due to the vital role of AD in circular economy, commercial AD biogas reactors have been constructed worldwide (Tao *et al*. 2020). However, the process of anaerobic digestion is very complicated. It requires the cooperation of various microorganisms to complete the four stages of hydrolysis, acidogenesis,

acetogenesis, and methanogenesis to obtain the expected product (Gujer, and Zehnder 1983). The process is schematically presented in Fig.1. There are still many problems that have always affected the efficiency of the AD reactor, such as low methane production, unstable system operation, and long start-up time (Chen, Cheng, and Creamer 2008), which indicates that further study on the AD microbiome is urgent, especially studies on hitherto unknown microorganisms may bring a breakthrough in our understanding of the process. Different configurations of AD reactors will affect its stability and production efficiency, such as temperature and food (Kim, Ahn, and Speece 2002). According to the temperature, the 20 AD reactors analysed in this study can be divided into three types, mesophilic(37-42°C), hyper mesophilic (44°C) and thermophilic (52°C), and their numbers are 12, 3 and 5 respectively. Throughout the AD process, different microorganisms require different substrates and these substrate requirements can also be viewed as a food chain in the AD reactor. Therefore, understanding the microbial distribution and cooperation within and between different levels in the microbial food chain is of importance for steering and improvements of the efficiency of the AD digestion.

1.3. Co-occurrence network in microorganisms

In recent years, the use of network analysis to study the co-occurrence patterns of microbial communities in complex environments has become an important method to understand the potential relationship between microorganisms, especially in the field of human gut and soil microorganisms (Faust *et al.* 2012; Barberán *et al.* 2012). To complete a conventional co-occurrence network analysis, the first step is to use some professional bioinformatics tools, such as Qiime and DADAD2 (Caporaso *et al.*2010; Callahan *et al.* 2016), to analyse 16S rRNA data to obtain operational taxonomic units (OTU) or amplicon sequence variants (ASV) tables, which containing taxonomic information and the relative abundance of microorganisms. Then Pearson or Spearman can be used to calculate the R-value between OTUs or ASVs, and finally software such as Cytoscape or Gephi can be applied to realize the visualization of the co-occurrence network (Shannon *et al.* 2003; Bastian, Heymann, and Jacomy 2009). The network is composed of many nodes and edges. Usually, nodes are set to represent different species and edges represent the correlation between nodes, the size of node represents its importance in the network and the width of edge represents the strength of the correlation. The network can also be painted according to different modules to show which nodes are more directly connected.

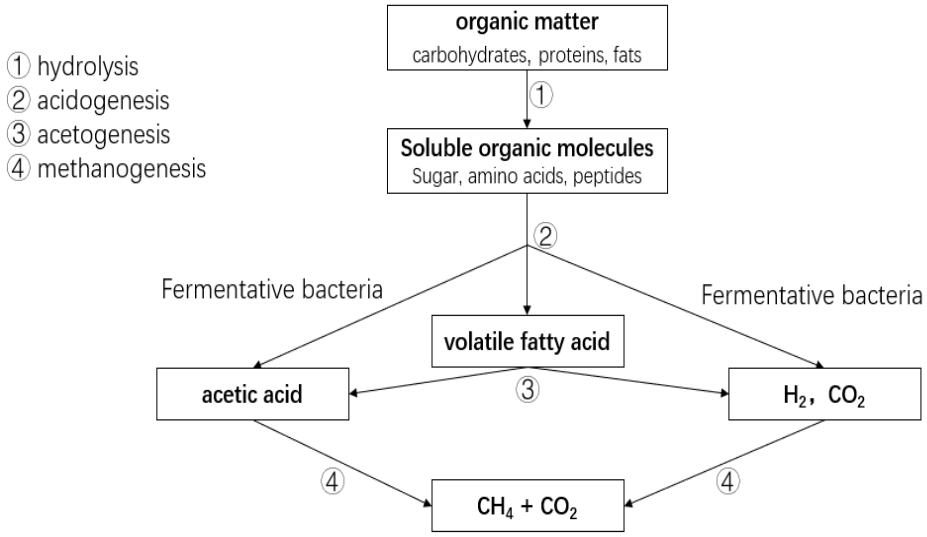

Figure 1. The whole process of AD system. Organic waste is converted into methane and Carbon Dioxide(CO_2) in a process.

2. Aim

In this study, metagenome-assembled genomes (MAGs) retrieved from metagenomics data of 20 AD biogas reactors and the respective Illumina sequencing reads were used to create an abundance table from which co-occurrence networks were build. The co-occurrence network also was done by 16S rRNA sequencing data to make a comparison between these two different sequence approaches. From that co-occurrence networks, we particularly wanted to gain insights into the importance and function of those microorganisms (MAGs) which were classified as unknown. These unknowns may represent potential microorganisms which cannot be cultivated but take over important function in the microbial food chain.

3. Method and material

3.1. Sampling and data

For all 20 AD reactors, 10 are from Sweden and 10 are from Germany, they are provided with food from different sources (organic household waste, slaughter house, manure, wastewater and green based such grass silage or maize) and run at different temperatures, mesophilic(37-42°C), hyper mesophilic(44°C) and thermophilic(52°C). Samples from all reactors have previously been taken and Deoxyribonucleic acid (DNA) prepared and sequenced using Illumina and Nanopore (Gloor *et al.* 2010; Branton *et al*. 2010). The sample preparations and sequencing were performed by previous study and from these sequencing data sets, in total 1122 bins (MAGs) with a completeness of at least 80% were recovered in previous work (Brandt *et al*. 2019). 16S rRNA amplicon sequencing data of all reactors were also analyzed and provided by my supervisor Bettina Müller.

3.2. Preparation of OTU table

Due to the lack of standardized bioinformatics tools, the generation of OTU tables in metagenomic data requires multiple steps. Referring to the OTU table production process of 16S sequencing data, all steps can be summarized as cluster, taxonomic classification and calculation of abundance. Before the present project started the taxonomic classification was already done by upstream experiment, in which the data of 20 reactors was sequenced by Illumina and Nanopore, and contigs after binning were classified to the strain level by Genome Taxonomy Database (Brandt *et al*. 2019). However, the OTU table usually needs to cluster the data to the species level, because the information about the deeper level in most databases is limited and the deeper classification level may cause too detailed classification and increase the difficulty of analyzing the network. Considering this situation, it is feasible to cluster the data back to the species level, and the simplest method is to find all strains belonging to the same species.

Because the taxonomic classification information of data was already available, it seemed feasible to start directly from the classification information. However, the data contains many unknown microorganisms, which do not have detailed classification information. The results obtained by relying only on classification information are not credible. Therefore, another method that can compare the similarity of two bins and suitable for unknown microorganisms is needed as verification.

Regarding different methods, Sourmash is an ultra-fast, lightweight nucleic acid search and comparison software, which can quickly analyze K-mer from DNA sequences by using min-wise independent permutations locality sensitive hashing scheme (Minhash) method to compare samples (Brown, and Irber 2016). The key to this method is the comparison of K-mer from different sequences, K stands for a number that can be artificially defined, and mer refers to the meaning of monomer unit in molecular biology, which usually be used in nucleic acid sequences, representing nt or bp. The k-mer refers to dividing the nucleic acid sequence into a string of k bases, that is, iteratively selecting a sequence of K bases from a continuous nucleic acid sequence, greater K means higher specificity of string. Usually, when k is 51 means compare two sequences on the species level.

Considering the needs of different software, all bioinformatic analyses are completed on the server. After getting the access to the server from the administrator, establishing the environment required to run various software is the first task. Miniconda2 was chosen to solve the software installation and operating environment problems, Miniconda is a lightweight alternative to Anaconda (Analytic 2016), which can directly install the required packages through simple commands and automatically check and install the lacked environment in the system.

When Sourmash (version 3.2.2) is installed, the input file must be in fasta format. Sourmash will first generate an independent sig file for each sample, and then obtain the result based on the comparison of the sig files. Illumina sequencing data was used because of its high accuracy. In the first attempt, all bins were put in and compared together. However, too many samples lead to pdf format results which cannot be clearly interpreted. In the second attempt, the scheme was adjusted to compare the data of 20 reactors in pairs and the result was improved, but this greatly increased the workload. The next step was to organize the matching results into a excel file and establish OTU in species level. For this step, no suitable automation software has been found to accomplish this task, and the matching information can only be put into the excel file by the most basic method, which is slow process. Writing some novel code may make this operation simple, but my level of expertise is not enough to complete this task. While entering matching information, it is necessary to use classification information for comparison to ensure the correctness of the results. Finally, 458 different species were obtained from 1122 bins. After a

week of data processing, the cluster and classification information were ready, leaving only the calculation of relative abundance.

Usually few programs are specifically used to calculate the abundance of metagenomic data, but these are not suited for analyses performed in this project, so an alternative strategy was necessary to apply. At the beginning, I tried to express the abundance by comparing the number of contigs in different bins which didn´t work. Because considering the sequencing depth and coverage, there is no guarantee that all fragments will have the same amount of replication, because the sequencing depth of different region in sequence cannot be same. After the discussion with my supervisors, mapping all bins back to their reactor data looks feasible.

Mapping is usually used for quantitative analysis of gene expression, which refers to comparing the reads produced by sequencing to the reference genome. Mapping results can provide information such as the position and matching quality of the read on the reference genome or other reference sequences (Schbath *et al.* 2012). The number of reads that match the reactor data in each bin can be obtained and used to represent abundance.

Three bioinformatic tools were found for mapping, bowtie2, metaphlan2 and salmon (Langmead, and Salzberg 2012; Segata *et al.* 2012; Patro *et al.* 2017). Because of the experience of using bowtie2 for analysis, this tool was first used. Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. Bowtie2 uses the FM (based on the Burrows-Wheeler Transform or BWT) index to index the genome for reducing memory usage (Langmead, and Salzberg 2012). When bowtie2 (version 2.3.5.1) was installed, the first step was building a library for each bin where the input file format requirement was fasta. For the reference genome data, it should be double-end sequencing data, meaning that Nanopore sequencing data is not suited and therefor Illumina data was used. Bowtie2 provides result in Sam format, which is a universal format for mapping result, and it contains the rating and location of the mapping. However, due to the lack of standardized statistical data, it was impossible to obtain convincing results to directly use the number of matching reads to represent abundance. To avoid this problem, the remaining two tools, metaphlan2 and salmon where tested. The analyses showed both tools could provide results that directly represent the abundance.

Metaphlan2 (version 2.7.6) was the second attempt for mapping. This tool is well suited for analyzing the composition of microbial communities, and it is very convenient in metagenomic research. Only one command is needed to obtain microorganisms species abundance information and it also has its own script for further statistics and visualization (Segata *et al.* 2012). The installation is very convenient, and the results are easily obtained. Furthermore, the results are in the format of an excel file including all species found and the abundance of them are

obtained. However, the problem is that it does not calculate the abundance according to each bin but classifies the species and calculates the abundance for each reactor, which is unexpected. Because the original classification information from previous experiment is reliable, the obtained results needed to be matched with the original classification information. The results where that they do not match, not even for a specific reactor. Some species that are impossible to exist in a reactor were identified. The reason for this may be that metaphlan2 can only use specific given databases, which will lead to errors in the identification of specific species.

Salmon is an alignment-free transcript quantification tool, which does not need to compare reads to a genome. Salmon divides the samples into k-mer, and then calculates the relative abundance of the samples by the number of k-mer matching the reference genome (Patro *et al.* 2017). Although it is mainly for the analysis of RNA sequences, its working principle is also applicable to the analyses of DNA sequences. The first step in using salmon (version 1.1.0) is to index each sample. Considering the workload, all bins from the same reactor were merged into one file and then matching those files to the genomes. Salmon will generate a sub-directory for the results of each sample, which contains quantitative results, running reports and some other information about the sample. Information about the quantitative statistics of the sample is stored in a tsv format file. The name of each contigs, their length, effective length, transcripts per million (TPM) and matching number of reads are included in this file. TPM is a standardized method to calculate abundance. The method calculates for a million RNA molecules in a sample, how many are from the reference transcript, and in the calculation process, the sequence length and sequencing depth are normalized, which is very convenient for comparisons within a single sample. When getting the relative abundance of contigs in all bins, calculation of abundance for each bin is needed. Since each bin contains an abundant of contigs, it seems that to represent the abundance of bins by calculating the sum or average of contigs abundance is two feasible methods. After discussion with my supervisors, the method of summation is found to have drawbacks, because there may be an overlap between different contigs in the bin, which will lead to a higher relatively abundance for each bin, and the average can minimize the impact of this situation.

3.3. Statistical and network analysis

The construction of co-occurrence networks is mostly based on correlation inference. Common methods of correlation inference include Pearson, Spearman and other methods (Artusi, Verderio and Marubini 2002). The Spearman method was selected because it is best to choose Pearson only when the dataset meets some certain conditions such as normality, homogeneity of variance etc. Although the

credibility of results calculated by Spearman is lower than that of Pearson, this problem can be solved by controlling the P-value and R-value selection intensity. The program Psych (version 1.9.12.31) was applied to calculate the R value (Revelle 2019). The program contains, an R package that can use different statistical methods to calculate correlation and it is applicable when the data volume is small. It can also provide a matrix for R and p values as input files for subsequent network analysis. The program iGraph (version 0.8.1) was used to finish the network analysis (Csárdi and Nepusz 2006). This program can realize various operations on the network according to the needs of the users, such as coloring the network according to different modules, coloring according to the classification information and obtaining the attributes of the network. However, when the number of nodes is too large, it cannot provide clear pictures under the basic process, and a more advanced R level is required. Due to lack of relevant experience, other software had to be used.

Cytoscape and Gephi were quite popular in network analysis, they are all open source software used to analyze complex networks and can provide network attributes and adjust the network according to user needs (Shannon *et al.* 2003; Bastian *et al*. 2009). Cytoscape (version 3.7.2) accepts input files in multiple formats and has examples of each format file. The most common input files are node lists and edge lists in csv format, and existing network data can also be used directly. However, due to the lack of node lists and edge lists, Cytoscape could not be used. Gephi (version 0.9.2) has the advantage that it can directly generate the network through the r value matrix, and then output the node list and edge list from the network, which solves this problem. Additional information can be added to the node list and edge list to improve and adjust the network, such as adding species information for classification, setting labels and characteristics of edges.

3.4. Annotation of selected co-occurrence cluster

The program Prokka (version 1.12) was used for annotation (Seemann 2014). In addition of annotation of prokaryotic genomes, this program can be used for gene identification, functional annotation and annotation file generation where multiple methods and databases are used to annotate gene functions in sequence. The input file must be in fasta format but results in 10 different formats can be produced. Blast (version 2.10.0) was also used for alignment to find potential link between metagenomics data and 16S data (Altschul *et al*. 1990).

4. Results and discussion

4.1. Co-occurrence network analysis

According to the different choice of type of reactors and p value, 4 networks were generated, all reactors ($p<0.05$) and only mesophilic reactors ($p<0.05$, $p<0.01$, $p<0.001$) respectively, the R-value of all the networks in this study is ($R>0.6$). The same networks based on 16S sequencing data were also generated for comparison. Because the number of reactors at the other two temperatures (thermophilic and hyper-mesophilic) are not enough to give statistical data, the network of all reactor ($p<0.05$) was used to compare with the network of mesophilic reactors ($p<0.05$). This to show all possible associations under a low selection intensity. Although it increased the chance of accidental occurrence, it is easier to find some neglected associations. The three mesophilic networks with different p-values were to show the impact of changes in p-values on associations and to find the associations with the highest confidence.

4.1.1. Threshold of R-value

Different choice of threshold has a significant effect on the network, but here only the results of different P-values were discussed in this study. However, there are many other choices of data threshold that are very important to construct a reliable network. Another aspect to consider is the R value. When making the OTU table, the matrixes of R and P values were generated at the same time, and then the final correlations were filtered according to the threshold. In this study, all networks have $R>0.6$. The reason for setting 0.6 as the threshold is that 0.5 and 0.7 are generally considered to represent the genus level and species level, respectively. OTU table in this study was annotated to species level. Using 0.5 is obviously too low and may cause more low-intensity correlations. Considering the annotation ability of 16S data at the species level and in order to retain more correlation, 0.6 was finally selected.

4.1.2. Threshold of relative abundance

Another important threshold is relative abundance. In this study, all species with a relative abundance less than 0.1% were removed, because previous studies suggest that relative abundance greater than 0.1% can be used to evaluate the importance of a particular species in the entire community (Tao *et al*. 2020). This also explains why the 16S data included more species, but with fewer nodes appearing in the network. However, when the information given in the network is very limited, lowering the threshold of relative abundance will help to get more correlations, but it may also cause errors. The problem can be solved by adjusting the R and P values. In the network analysis, the selection of these three thresholds (write these in the text) are very important, but there is no clear solution to use a standard choice in advance. All threshold selections must be adapted to the biological problems to be solved.

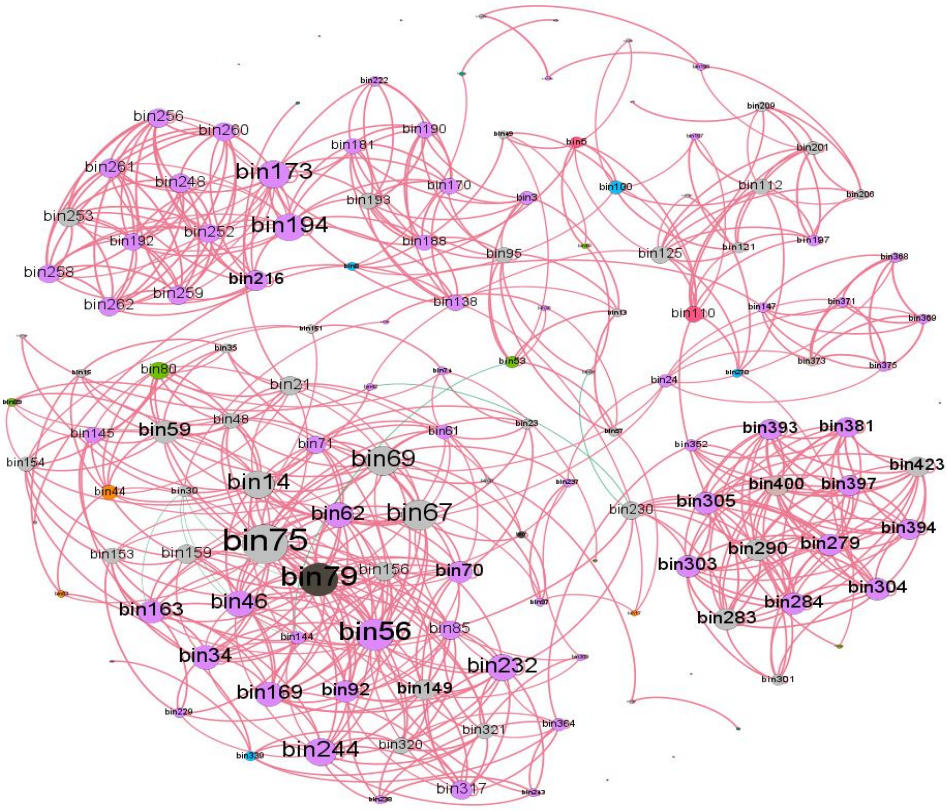

Figure 2. The network of all reactors (P-values <0.05). The size of nodes means the importance degree of the node in the network calculated by the software and the nodes were colored according to taxonomic classification information. The purple nodes mean unknown microorganisms. The red edges mean positive correlations and green edges mean negative correlations.

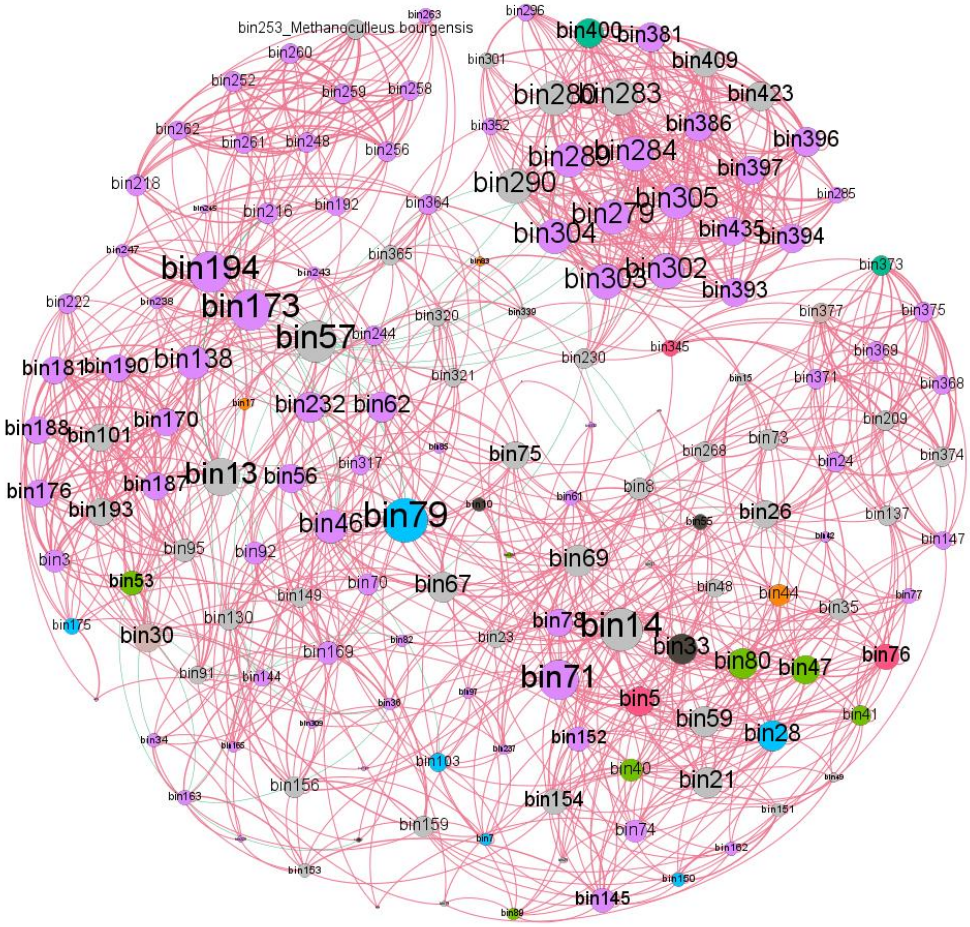

Figure 3. The network of mesophilic reactors (P-values <0.05). The size of nodes means the importance degree of the node in the network calculated by the software and the nodes were colored according to taxonomic classification information. The purple nodes mean unknown microorganisms. The red edges mean positive correlations and green edges mean negative correlations

4.2. Changes in the network under specific temperature conditions

After analyzing p-value using the thresholds discussed in section 4.1.2 some network containing interesting nodes and edges was obtained. The network for all reactors ($p<0.05$) has 142 nodes and 608 edges (Figure 2) and for mesophilic reactors ($p<0.05$) has 159 nodes and 1201 edges (Figure 3). It is easy to find that the mesophilic network has more nodes and edges. More nodes indicate that in the mesophilic environment, the abundance of bins is relatively more average than that

of all reactors. Under mesophilic conditions, more edges indicate that the correlation between nodes in a single temperature is more frequent. This because network of all reactors contains nodes from three different temperatures, which also reduced the correlation between them. Due to the insufficient number of reactors at the other two temperatures (hyper mesophilic and thermophilic), convincing statistical R-value result was unavailable. Hence, the following analysis will focus on the network of mesophilic reactors.

4.3. Potential co-occurrence cluster in AD reactor

More clear clusters can be found in mesophilic network and each cluster will be connected to each other through some nodes as bridges to form the entire network, Because the selection strength of the p value is not high, there are almost no independent clusters. After checking the classification information, bin253 (known as *Methanoculleus bourgensis* according to the taxonomic information) was found to be the only one of the three methanogens that appeared in a cluster. The function of methanogens is central and of high importance as they perform the last step in the anaerobic food chain, the formation of methane. Therefore, it is necessary to explore this cluster. The cluster can be found on the top of Figure 3 and it is connected to a cluster under it through two nodes bin173 and bin194 as a bridge. The distribution of these two clusters can be found through the connection situation of bin173 and bin 194 (Figure 4). In the partial graph of mesophilic network, two clusters can be clearly distinguished, bin253 is included in the upper cluster and all other nodes in this cluster are unknown, and in the cluster below, more known nodes can be found. Nodes with negative correlations to bin173 and bin194 are form other clusters. These two clusters might be two connected levels in the food chain system, but further analysis of the involved bins is needed to verify this hypothesis.

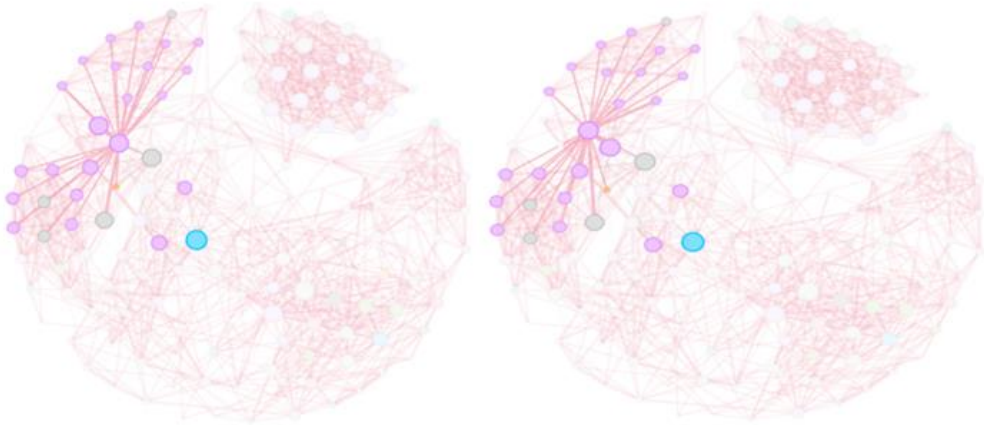

Figure 4. The connection of bin173 (left) and bin194 (right) reveals the distribution of two clusters. The purple nodes mean unknown microorganisms. The red edges mean positive correlations and green edges mean negative correlations.

4.4. The influence of different p-values on the network

When the p-value was changed, the result showed that the mesophilic network also changed significantly. Compared with the network (p-values<0.05), the network (p-values<0.01) has fewer correlations, 682 edges (Figure 5). After removing some lower confidence edges, the entire network becomes clearer, and each cluster is also easier to identify, where at least five clusters can be found. However, the previously mentioned bin173 and bin194 cannot be found, which indicates that the function of these two bins might not be a necessary way to connect the two levels of the food chain in mesophilic AD reactors, at least they disappeared under high confidence conditions. Nevertheless, it is still necessary to compare the results of different p-values. When the p-value selection strength is lower, the results closer to the original network are found. Even if there are some accidental edges, we can still obtain some overall information, which is not available under high p-value selection intensity. It is worth noting that the cluster containing bin253 can still be found, which shows that even if the selection intensity of p-value is increased, this cluster still plays a key role in the network.

When the threshold of p-value is reduced to 0.001, a very high selection intensity, the network changed a lot (Figure 6). There are only 318 edges left, and the entire network becomes clearer. It can be found that five clusters are distributed on the network and they have no correlation with other nodes, which means that these five clusters may be stable and might exist independently. Conclusively, these five clusters may play a crucial role in mesophilic AD reactors. The cluster containing bin253 still exists in the network, and all the nodes in this cluster have a positive correlation with bin253, which indicates that they might be cooperative and work together to achieve specific functions. Methanogens mainly participate in the final step of the AD process, where they form methane from H_2 and CO_2 and from acetate. MAB1, a strain of *M. bourgensis*, was identified as the hydrogenotrophic partner of mesophilic acetate-oxidizing bacteria. This syntrophic relationship is closed to the thermodynamic equilibrium and vital for ammonia-rich engineered biogas processes (Manzoor *et al.* 2016).

A quick screen of the annotation results revealed that the metagenome-assembled genomes Bin 260, 252, 262, 248, and 256 encode a very limited number of transport systems for both carbohydrates and peptides. In accordance to that, none of them appears to encode features needed for catabolite repression. This indicate a very specialized metabolism as observed for syntrophic microorganisms. Additionally, all these Bins harbor genes coding for soluble Fe-Fe hydrogenases. Fe-Fe hydrogenases are typically found in organisms, which using protons as terminal electron acceptor such as syntrophic acetate or propionate oxidizers or as

electron sink such as acetogens. Both requires the presence of hydrogenotrophic methanogens to be efficient and in the case of syntrophic interactions, to be thermodynamically possible. On opposite, the metagenome-assembled genomes Bins 258, 259 and 261 harbor a significant number of transport systems predict to transport small oligo carbohydrates and carbohydrates, many predicted as phosphotransferase systems (PTS), what points to global regulation by catabolite repression. These bins also encode soluble Fe-Fe hydrogenases indicating that the presence of hydrogenotrophic methanogens might be beneficial for the energy metabolism of these microorganisms by redirecting electrons towards hydrogen production.

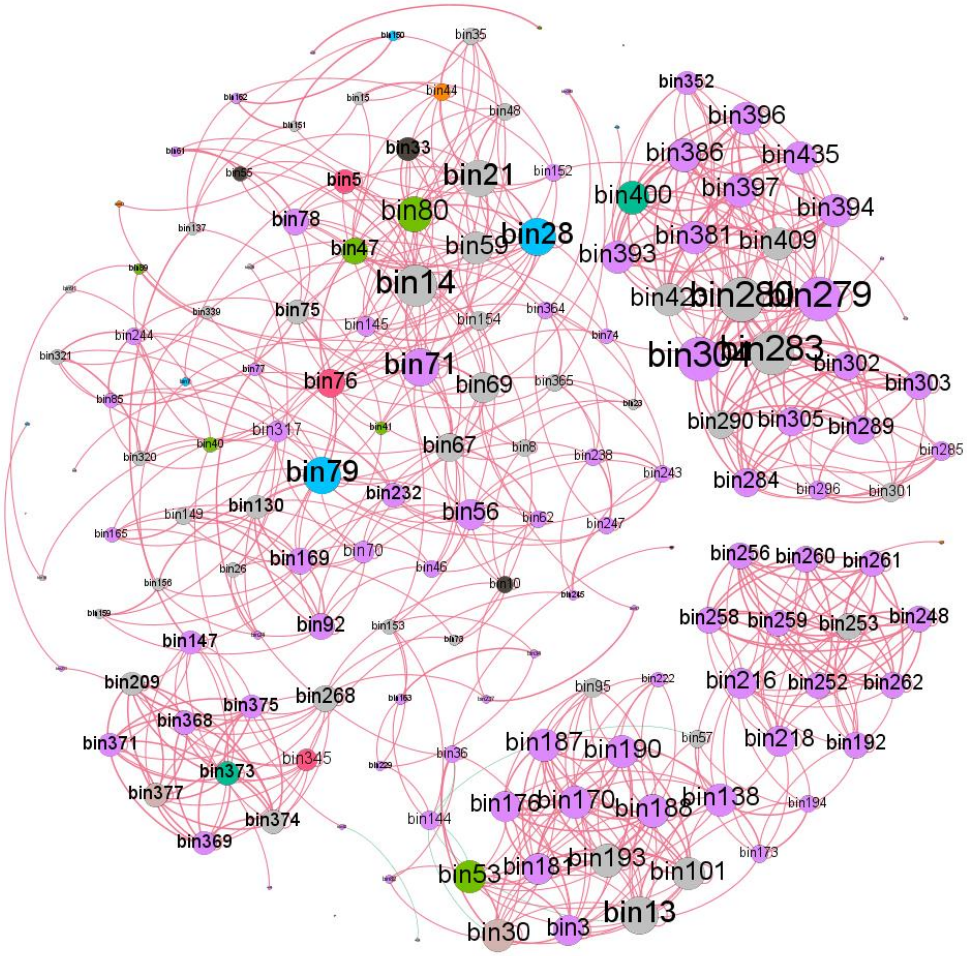

Figure 5. The network of mesophilic reactors (P-values <0.01). The size of nodes means the importance degree of the node in the network calculated by the software and the nodes were colored according to taxonomic classification information. The purple nodes mean unknown microorganisms. The red edges mean positive correlations and green edges mean negative correlations.

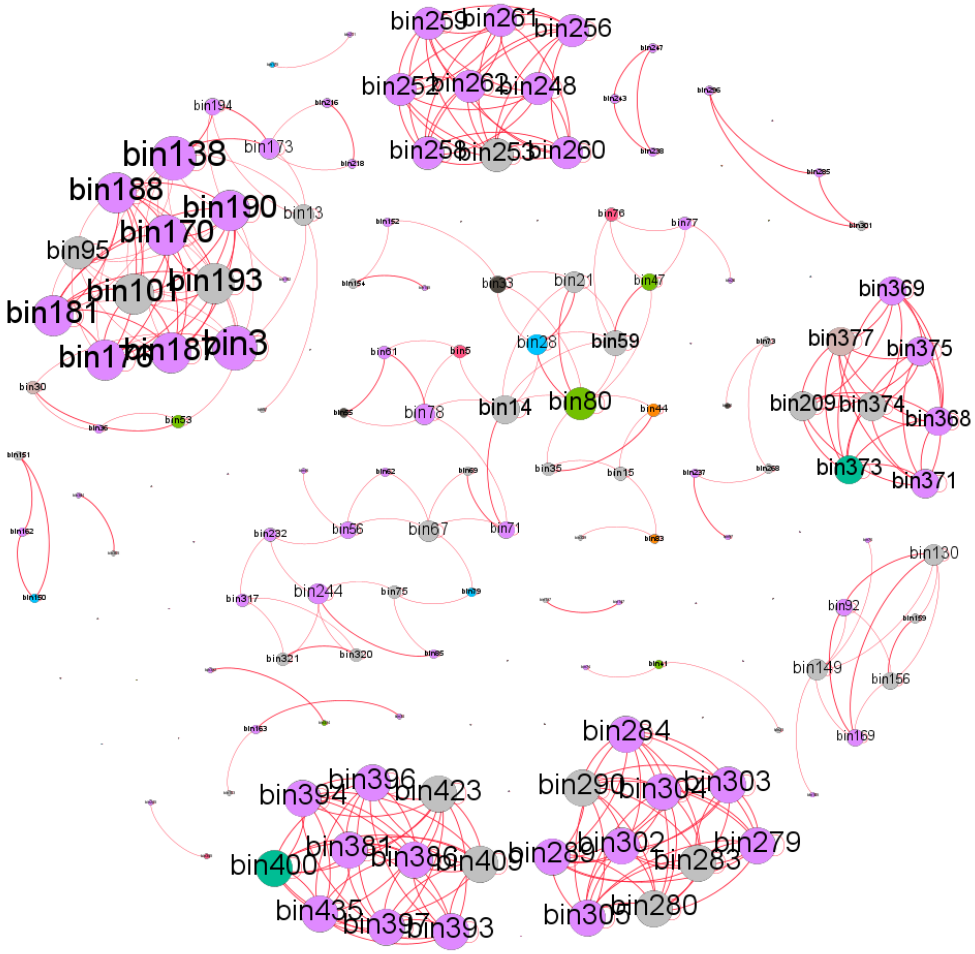

Figure 6. The network of mesophilic reactors (P-values <0.001). The size of nodes means the importance degree of the node in the network calculated by the software and the nodes were colored according to taxonomic classification information. The purple nodes mean unknown microorganisms. The red edges mean positive correlations and green edges mean negative correlations.

4.5. Network comparison of 16S data and metagenomic data

In all reactors, 16S sequencing data found 551 species, which is more than that 458 species found in metagenomic data (reference to the 458 results), which suggests that there will be more correlations in the network. However, the 16S mesophilic network showed opposite results. The mesophilic network (p-value<0.05) of 16S

data has only 110 nodes and 872 edges and network (p-values<0.001) has 110 nodes and 123 edges (Figure 7), the edges numbers under both p-values are much less than that in metagenomic network. In 16S network (p-value<0.05), no cluster can be clearly identified, and the entire network looks very chaotic. For 16S network (p-value<0.001), only one cluster can be found. The information we can get from the 16S network is very limited, which is contrary to the more species obtained from the 16S data, while the metagenomic data obtained less species information but provided more network information. The reason for this may be the difference in their sequencing methods. 16S sequencing usually selects one or several of the 9 highly variable regions in the 16S rDNA gene for amplification and sequencing, while metagenome sequencing randomly breaks the entire genome and then perform amplification, assemble and sequence. When the classification information is annotated to the species level, from 16S sequencing, each highly variable region has a high specificity. Although some species may be very similar in these regions the specific fragments that can distinguish them may not be in the amplified region. Another difficulty is how to define species, usually two 16S sequences with a similarity greater than 97% are considered to belong to the same species, but many bioinformatics tools will categorize some small differences as the default fault tolerance during alignment, which will also reduce the accuracy of species information. In addition, it is assumed that many species are not recognized by the 16S primers currently used. The fact that the metagenomics approach in this project recovered more than 500 MAGs which could not be assigned to a previous reported taxon supports this assumption. Thus, the 16S data set apparently lack important information which results in poor cluster formation when building the co-occurrence network. Due to the time limitation of this project, this study did not compare the two data analysis methods at the genus level, but one can assume that the network provided by the two kinds of data at genus level will be closer.

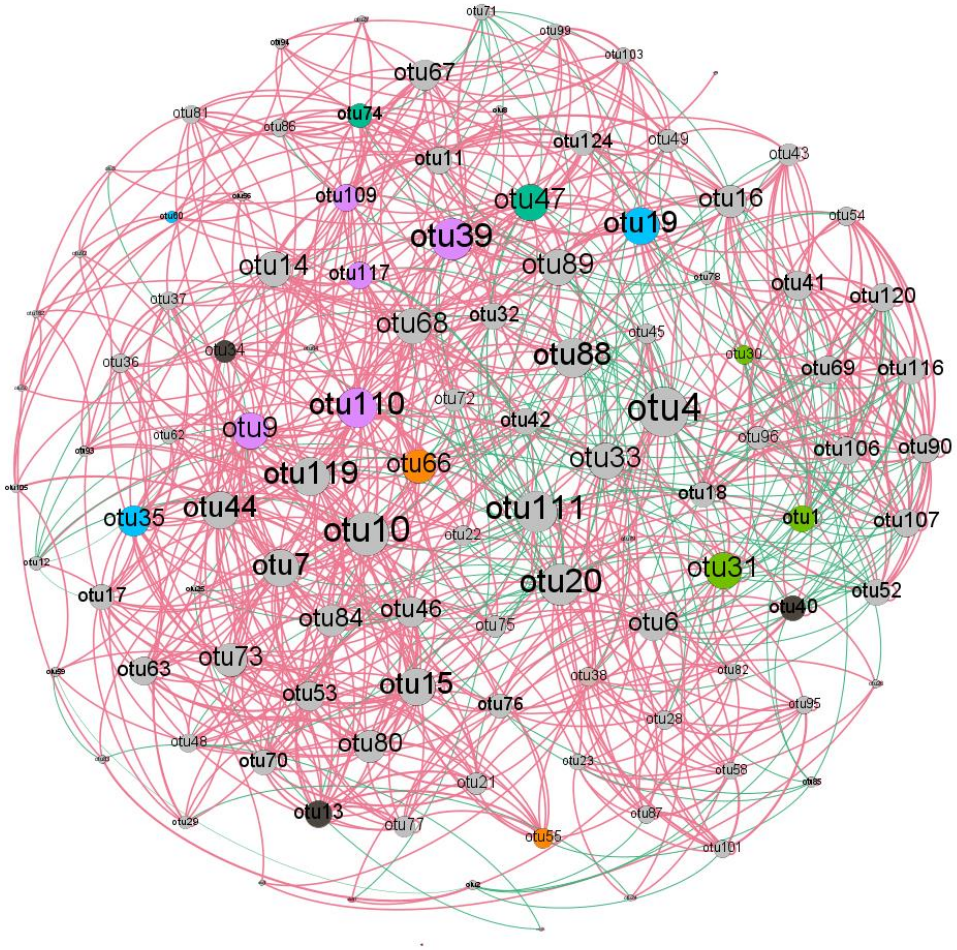

Figure 7. The network of 16S mesophilic network (P-value<0.05). The size of nodes means the importance degree of the node in the network calculated by the software and the nodes were colored according to taxonomic classification information. The red edges mean positive correlations and green edges mean negative correlations.

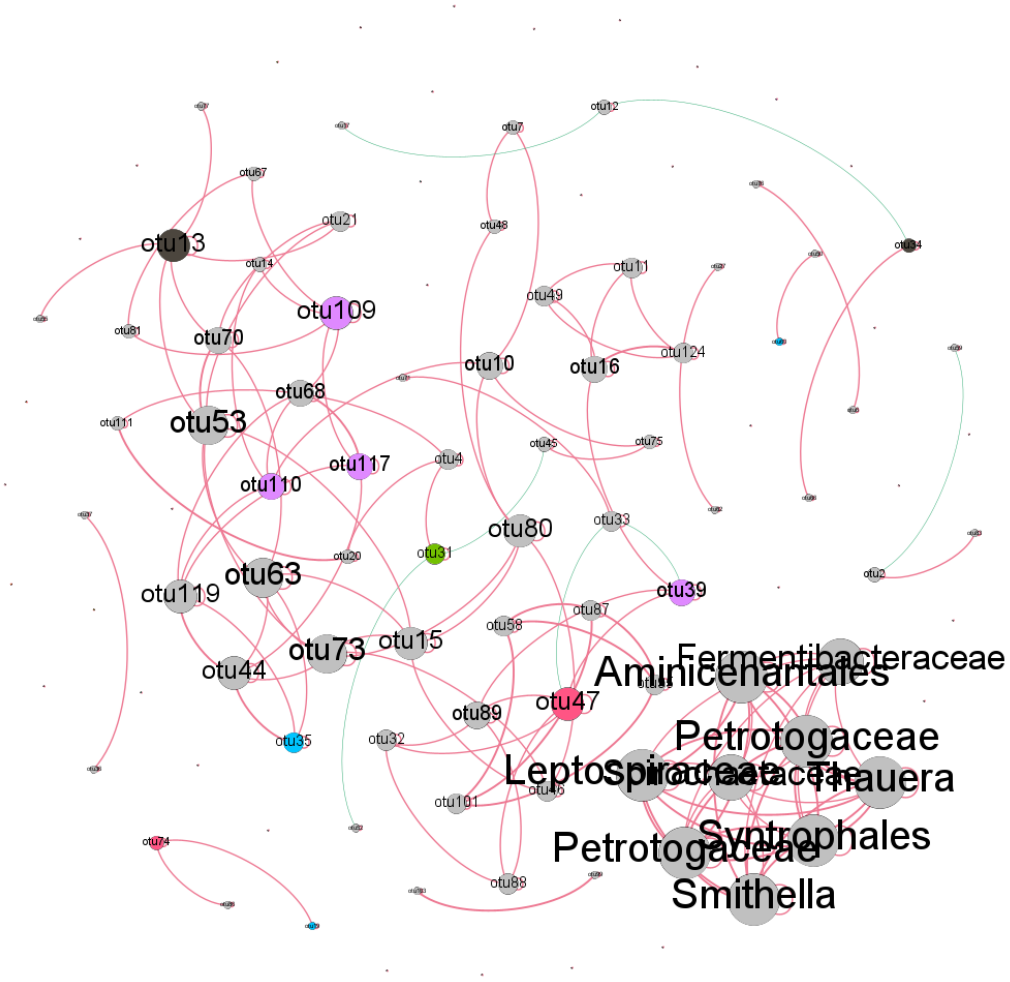

Figure 8. The network of 16S mesophilic network (p-value<0.001). The size of nodes means the importance degree of the node in the network calculated by the software and the nodes were colored according to taxonomic classification information. The red edges mean positive correlations and green edges mean negative correlations.

5. Conclusion and future outlook

A feasible method was designed for making OTU table and network analysis of metagenomics data. In the network, some potential clusters were found, and the cluster where bin253 is located may belong to methanogenesis level, but further metabolic functions and pathway analysis are needed. Also found was that the bins classified as “unknown” form clearly supported clusters. This indicates that these microorganisms are heavily dependent on the interactions with other microorganisms. This makes perfect sense; such microorganisms are difficult to cultivate and characterize thereby usually referred to as “unknown”. Metagenomics in combination with co-occurrence network analysis enables both the discovery of these microorganisms and the elucidation of their potential function within the microbial community.

Due to the limited time of this study further development of the methods was restricted. In the future improvements can be made for example in the process of making the OTU table, the mapping results of Nanopore data can be added this to improve the abundance information. Although Nanopore data usually has low accuracy, it has a better sequencing depth. When trying a new method, only one kind of data will often cause errors, why it is important to combine several methods to do more comparisons and verifications. Another improvement that is urgent is to develop faster methods, the whole bioinformatics pipeline currently applied was complex and time-consuming. When different networks need to be generated, even if only the p-value is changed, a lot of information on the node list needs to be manually reprocessed, which is a particularly simple but time-consuming work. Comparing the mapping results is also problematic due to the lack of matrix files, which means that matching the results needs to be done manually. This takes at least 20 hours to solve manually but could be much faster by automation. In the future, more automated software can be designed where the entire process is fully automated, and the corresponding network can be obtained by simply adjusting the threshold. It should be feasible to implement such software according to the current technical level.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.

Analytics, C. (2016). Anaconda software distribution. *Computer software*. Vers, 2-2.

Artusi, R., Verderio, P., & Marubini, E. (2002). Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval. *The International journal of biological markers*, 17(2), 148-151.

Barberán, A., Bates, S. T., Casamayor, E. O., & Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME journal*, 6(2), 343-351.

Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: an open source software for exploring and manipulating networks. *In Third international AAAI conference on weblogs and social media.*

Brandt, C., Bongcam-Rudloff, E., & Müller, B. (2019). Abundance tracking by long-read nanopore sequencing of complex microbial communities in samples from 20 different biogas/wastewater plants.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., ... & Jovanovich, S. B. (2010). The potential and challenges of nanopore sequencing. In Nanoscience and technology: A collection of reviews from *Nature Journals* (pp. 261-268).

Brown, C., & Irber, L. (2016). sourmash: a library for MinHash sketching of DNA. *Journal of Open Source Software*, 1(5), 27.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, 13(7), 581.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... & Huttley, G. A. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5), 335.

Chen, Y., Cheng, J.J. & Creamer, K.S. (2008). Inhibition of anaerobic digestion process: A review. *Bioresource Technology*, vol. 99 (10), pp. 4044–4064 Elsevier Ltd.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, complex systems, 1695*(5), 1-9.

Daniel, R. (2005). The metagenomics of soil. *Nature Reviews Microbiology*, 3(6), 470-478.

Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C. & Ouzounis, C.A. (2012). Microbial Co-occurrence Relationships in the Human Microbiome (Human Microbiome Co-occurrence Relationships). *PLoS Computational Biology*, vol. 8 (7), p. e1002606 San Francisco, USA: Public Library of Science.

Gloor, G. B., Hummelen, R., Macklaim, J. M., Dickson, R. J., Fernandes, A. D., MacPhee, R., & Reid, G. (2010). Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products. *PloS one*, 5(10).

Gujer, W., & Zehnder, A. J. (1983). Conversion processes in anaerobic digestion. *Water science and technology*, *15*(8-9), 127-167.

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, vol. 68 (4), pp. 669–66985

Hugenholtz, P., Goebel, B.M. & Pace, N.R. (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. *Journal of Bacteriology*, vol. 180 (24), pp. 6793–6793

Kim, M., Ahn, Y.-H. & Speece, R.. (2002). Comparative process stability and efficiency of anaerobic digestion; mesophilic vs. thermophilic. *Water Research*, vol. 36 (17), pp. 4369–4385 Elsevier Ltd.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), 357.

Manzoor, S., Schnürer, A., Bongcam-Rudloff, E., & Müller, B. (2016). Complete genome sequence of Methanoculleus bourgensis strain MAB1, the syntrophic partner of mesophilic acetate-oxidising bacteria (SAOB). *Standards in Genomic Sciences*, 11(1). https://doi.org/10.1186/s40793-016-0199-x

Patro, R., Duggal, G., Love, M., Irizarry, R., & Kingsford, C. (2017). Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nature Methods*, 14(4), 417–419. https://doi.org/10.1038/nmeth.4197

Revelle, W. (2014). psych: Procedures for psychological, psychometric, and personality research. *Northwestern University, Evanston, Illinois*, *165*, 1-10.

Riesenfeld, C. S., Schloss, P. D., & Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.*, 38, 525-552.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., & Gibrat, J. (2012). Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of Computational Biology : a Journal of Computational Molecular Cell Biology*, 19(6), 796–813. https://doi.org/10.1089/cmb.2012.0022

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8), 811.

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Tao, Y., Ersahin, M.E., Ghasimi, S.M.D., Ozgun, H., Wang, H., Zhang, X., Guo, M., Yang, Y., Stuckey, D.C. & van Lier, J.B. (2020). Biogas productivity of anaerobic digestion process is governed by a core bacterial microbiota. *Chemical Engineering Journal*, vol. 380, pp. urn:issn:1385–8947

Acknowledgements

I would thank my supervisors Erik Bongcam-Rudloff and Bettina Müller. Erik is my beacon in bioinformatics and after his course I really ignited my interest in bioinformatics. Bettina's encouragement and support helped me to have the courage to move forward when facing difficulties.

I would also thank Hadrien Gourlé and Juliette Hayer for their help in bioinformatics.

Then I want to thank Renaud Van Damme and Xavier Goux, their patient answers and suggestions allowed me to complete this journey more firmly.

At last, I want to thank my mom, she supports me as always, and gives me the motivation to overcome all difficulties.

Now, we are going through a difficult period, and thank all the people who have helped me so that I can successfully complete the project.