# Development Of Database Tools to Manage Data on Complex Microbial Population in Wastewater Treatment Processes

Student ID 47-096824
Student Name Purnika Damindi RANASINGHE
Supervisor Assoc. Prof. Hiroyasu SATOH

## 1.Introduction

Activated sludge processes are the most widely used biological wastewater treatment methods. Activated sludge is essentially mixed culture of microorganisms attached together to form what are called "flocs". Flocs have higher density than water, and can easily be separated from treated water gravimetrically (Mino, 2000).

There are different microorganisms with different functions in activated sludge. Some of them are apparently helpful for treatment, such as polyphosphate accumulating organisms and nitrifying organisms, while others such as filamentous microorganisms are detrimental for treatment. Thus, understanding the microbial diversity and their population dynamics has become a major concern. In late 1960s-1980's conventional techniques like, cultivation-dependent plate counts and most probable number (MPN) methods were practiced, but due to their limitations molecular methods were introduced in late 1990's. Fluorescent in situ hybridization (FISH) has been introduced to observe cells hybridized with fluorescently labeled oligo-nucleotide probes for in situ identification of microbial species. Profiling methods of whole microbial community such as terminal-restriction fragment length polymorphism (T-RFLP) and denaturing gradient gel electrophoresis (DGGE) in combination with polymerase chain reaction (PCR) targeted at 16S rRNA gene are also widely used.

And today, pyrosequencing method is expected to make a breakthrough in analyzing environmental microbial samples including activated sludge, as the method will provide both taxonomic information and their abundances. Roche 454 pyrosequencer series has a power to yield 1 million reads each with 400bp in one run. In one run, up to typically 8 samples can be analyzed. But if it is arranged so that DNA fragments from each sample are labeled with unique barcode, samples can be mixed together, and then analyzed. Later, the origin of the samples can be assigned based on

the barcode sequence. Together with the development of sequence analysis methods, data handling methods are in rapid development. Ribosomal database project (RDP) is now providing useful tool for the analysis of 16S rRNA, including a set of tools for pyrosequencing data (pyrosequencing pipeline). Software like WATRES and QIIME (Quantitative Insight Into Microbial Ecology, Caporaso et al., 2010) have been developed as a single platform for complete pyrosequencing data analysis.

It is now expected that data from microbiological analysis work will increase in the near future. Here is a need to anyhow develop microbiological data handling tools that are helpful for researchers in wastewater science and engineering.

## 2.Objectives

The present study was conducted with following three objectives.

1. Establishing computational workflow for 16S rRNA pyrosequencing data analysis
2. Developing a "floc library", which is a compilation of morphological and molecular data from single flocs.
3. Developing reference sequence database for microbial groups that are thought to be representative or important species from different wastewater treatment processes

## 3. Materials, Methods, and Basic Concepts of Database Development

### 3.1 Preparation of Template DNA Sequences for Workflow Development

Activated sludge samples collected from different sources, including laboratory activated sludge reactors, urban and industrial wastewater treatment plants.

Samples stored at -80°C were thawed, diluted 20 times with Milli-Q water, sonicated by 250DA Advanced Digital Sonifier (Branson) with a special micro tip at an amplitude of 40% (20W) for 20 seconds. Then, PCR reaction was performed using barcoded universal primer pair 27f/519r which is targeted at a partial 16S rRNA gene (Lane, 1991).

Thermal cycles were programmed, 95°C for 600 seconds (Initial denaturation), 94°C for 30s, 55.3°C for 30s and 72°C for 30s for 30 cycles (denaturation, annealing and extension), 72°C for 600s (final extension) using Thermal Cycler Dice (Takara, Japan). PCR products were purified by QIAQuick purification kit and samples were submitted for 454 Titanium (Roche) Pyrosequencing. Pyrosequencing work was done by Center for Omics and Bioinformatics, Graduate School of Frontier Sciences, The University of Tokyo.

### 3.2 Development of Workflow to Process Pyrosequencing Data

In the first stage, data analysis was tried with reads from only two of the samples. Different data analysis methods such as RDP's Pipeline Process (Maidak et al., 1997) and ARB (Buchner et al., 2004) were combined and tried.

In the second stage, QIIME was introduced to data analysis workflow. QIIME sorts out reads to operational taxonomic units (OTUs), pick representative sequences for each OTU, assign taxonomy, and calculates phylogenetic tree.

In the third stage, OTUMAMi (Operational Taxonomic Unit Management and Mining, Satoh, 2011) was introduced and used in combination with QIIME. OTUMAMi is a workflow helper that generates commands for QIIME and is also a data organizer. The workflow was tested with reads from a larger set of sequences.

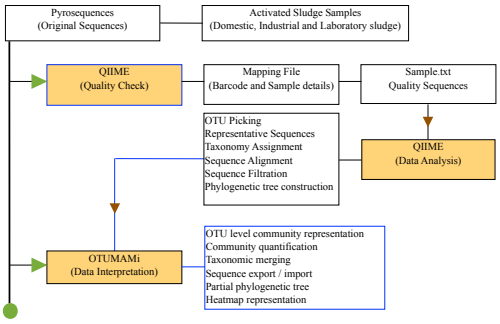


Fig.1 proposed workflow for pyrosequencing

### 3.3 Development of Floc Library

Activated sludge samples were collected at aerobic phase from four (A-D) laboratory scale Sequencing Batch Reactors (SBR) at different time intervals. Single flocs were isolated from activated sludge samples by observing under microscope. And their morphological characters including size, shape and color recorded with microphotographic images.

Individual floc samples were added with Milli-Q water or 50% ethanol to a volume of 1mL, and sonicated by 250DA Advanced Digital Sonifier (Branson) at amplitude of 30% (20W) for 20 seconds. And, Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) was performed using universal primer pairs (27f/ 519r) with PrimeScript® One Step RT-PCR Kit Ver.2 (Takara). Thermal programme was 50°C for 30 min, 94°C 2 min (Initial Denaturation), followed by 30 cycles of 94°C for 30 sec, 55.3°C for 30 sec, and 72°C for 30 sec, and final extension was done at 72°C for 10 min using Thermal Cycler Dice (Takara, Japan). The PCR products were digested by *HhaI* restriction enzyme and restriction fragment length polymorphism (RFLP) analysis was performed by Agilent BioAnalyzer with DNA 1000 Assay Kit. The RFLP data was utilized to compare microbial community in flocs.

Morphological characteristics and RFLP analysis data were imported into a Floc library developed by FileMaker Pro Advanced (ver.10.0).

### 3.4 Development of Reference Database

The reference database was developed as a storage of reads and taxonomic information for representative OTUs found from different activated samples. As OTUs stored in this database is expected to serve as "references", relatively small number of OTUs is thought to be selected and stored here. And preferably, the database should be accompanied by detailed information related to the reference OTUs.

Thus, the reference database was designed to have two functions. Firstly, it was designed so that uses can further select OTUs to be regarded as reference. Secondly, it was designed to so that uses can access more detailed information related to the OTUs.

A layout was developed to select OTUs for reference database. Then sequences for the selected OTUs were analysed by RDP's "Seq Match" to obtain best-match sequences in Genbank, or more exactly, accession numbers to the best-match sequences. The accession numbers were imported to the reference database, URL to the homepages for the best-match sequence was generated using the accession numbers, and the homepages were arranged to be displayed on the layout using the WebViewr function of FileMaker Pro.

## 4. Results and Discussion

### 4.1 Development of Workflow to Process Pyrosequencing Data

The combination of QIIME with OTUMAMi made pyrosequencing data analysis faster and easier than before. Data with 734976 reads obtained for all the activated sludge samples could be analyzed by the workflow within one day. In the workflow, manual handling of data was minimized, and even when it is needed, OTUMAMi gives comprehensive instructions. The manual data analysis and interpretation methods practiced prior to the development of OTUMAMi were discarded.

As shown in Fig. 1 pyrosequencing data was first processed by QIIME and the results were returned as several text files. These files were imported to OTUMAMi, and data from these files were re-organized to help users grasp the outcomes from pyrosequencing. Users can see the pyrosequencing results at different hierarchical levels from species level (or OTU level) as shown in Fig. 2. The OTUs are arranged vertically, and samples are arranged horizontally. The values in each cell are the fraction of the reads grouped to the OTU in the total number of reads for the sample. The fractions are displayed as heatmap: The color intensity of cells represent the value of the fraction: the higher the fraction, more intensified color is assigned.

In OTU-level heatmap, the OTUs are sorted in the order they appear in phylogenetic tree. A part of the OTUs can be selected, sequences for them are exported and re-analyzed by QIIME, phylogenetic tree is drawn by such programs as FigTree, and then, the tree can be combined with the OTU heatmap for the selected OTUs (Fig. 3).

### 4.2 Floc Library

Out of nearly 100 sludge flocs, 72 samples that were successful at RT-PCR, were selected. Floc library was developed as a collection of activated sludge flocs with different structure. Specially, shape, size and color. Pictures taken during microscope examination, RFLP patterns and correspondent restriction fragment lengths were also included to compare microbial communities (Fig. 4).

The developed floc library combines conventional morphological observations and molecular analyses data. Addition of new data to the library is also possible and users can compare the sludge morphology and microbial community diversity.


Fig.2 OTUs arrangement and community composition by OTUMAMi

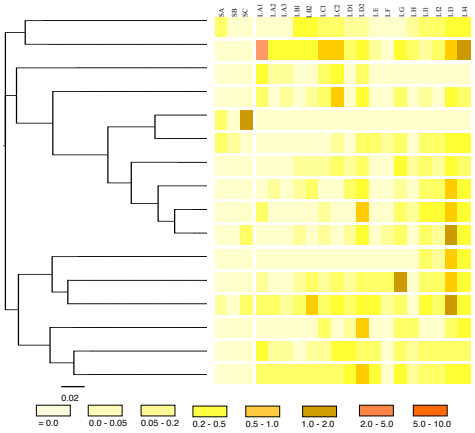


Fig.3 Combination of phylogenetic tree and heatmap structure

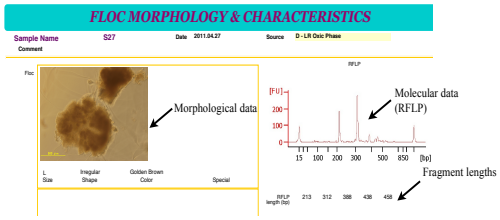


Fig. 4 Floc library record on morphological and molecular data

### 4.3 Reference Database

After the analysis of all pyrosequencing reads (734,976) obtained from all activated sludge samples, nearly 41,000 OTUs were generated. Selection criteria set to the fraction (>0.008) of community compositions and select the major microbial communities. The reference database provides the space to store OTUs records and allows selecting sequences of interests. Selection was further narrowed down to 430 OTUs which were thought to be representative OTUs and to cover all Phyla to Genus level in phylogeny. These OTUs were imported to the reference database. After the RDP analyses, collected data (text based) was imported to table "Reference Database" and presented by interface as in Fig. 5. The best matching sequences from Genbank annotation was presented and allows user to get familiar with related source of isolation, full or partial sequences in FASTA or Genbank formats, without BLAST analysis.

Thus, the reference database allows user to further select OTUs to be regarded as references or interests. And, composed with detailed information related to the OTUs.

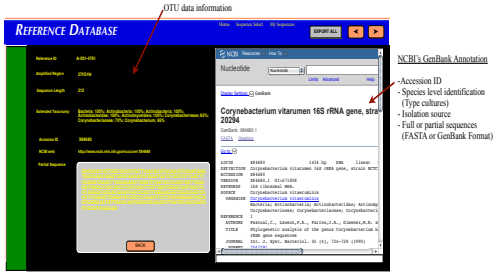


Fig.5 Reference database record with OTU data and Genbank annotation.

### 5. Conclusion

The present study, initiated to provide a database tools for the analysis of 16S rRNA pyrosequence data from activated sludge samples.

In the present study, three methods were tried, and finally, the workflow with OTUMAMi and QIIME was introduced as a reasonably effective method for data processing. Secondly, the floc library database was developed to store morphological and molecular information of flocs isolated from activated sludge. Thirdly, the reference database was developed. It helps constructing a database of OTUs which are important and thought to serve as references. Complete developments were highly tested using template pyrosequences obtained from different activated sludge samples. And, proved it's rapid data analysis skills and graphical data interpretation on a single platform.

### 6. References

Caporaso et al., (2010). Correspondence QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nature*, **7**, 335-336.

Maidak et al., (1997). The RDP (Ribosomal Database Project). *Nucleic acids Res*., **25**(1), 109-111.

Mino, T. (2000). Microbial selection of polyphosphate-accumulating bacteria in activated sludge wastewater treatment processes for enhanced biological phosphate removal. *Biochemistry. Biokhimica,* **65**(3), 341-348.

Buchner et al., (2004). ARB: a software environment for sequence data. Nucleic Acids Research, **32**(4), 1363-71.