2011年度　修　士　論　文

廃水処理プロセス中の複合微生物群集に関する情報を扱うためのデータベースツールの開発

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

Purnika Damindi RANASINGHE

プルニカ　ダミンディ　ラナシンハ

東京大学大学院新領域創成科学研究科
社会文化環境学専攻

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.

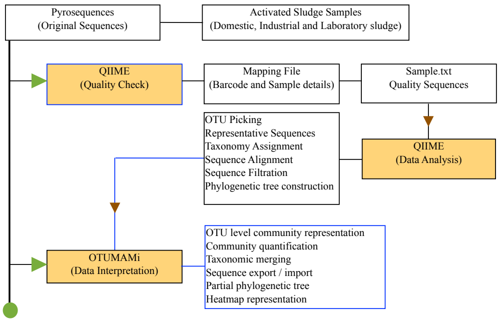


Fig.1 proposed workflow for pyrosequencing

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.

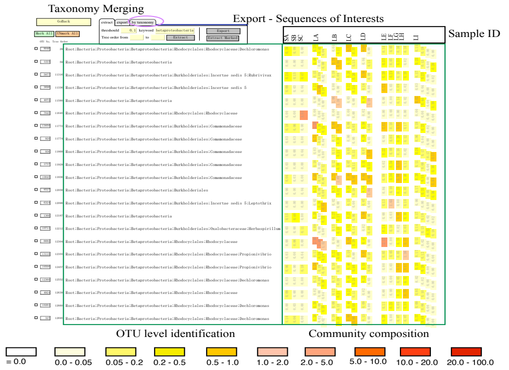


Fig.2 OTUs arrangement and community composition by OTUMAMi

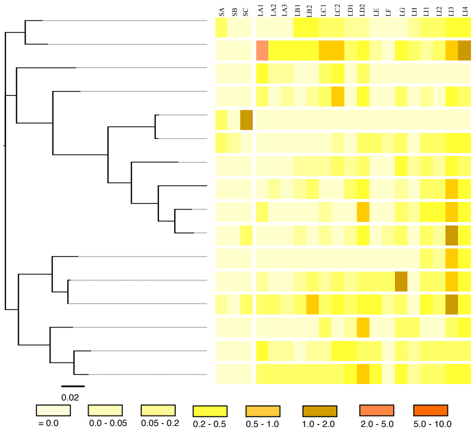
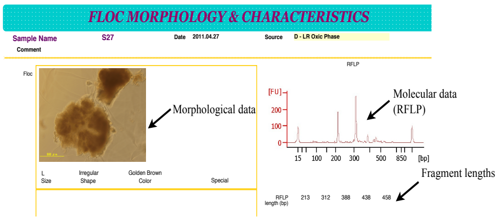


Fig.3 Combination of phylogenetic tree and heatmap structure



Sample Name S27 Date 2011.04.27 Source D - LR Oxic Phase

Comment

L Size | Irregular Shape | Golden Brown Color | Special

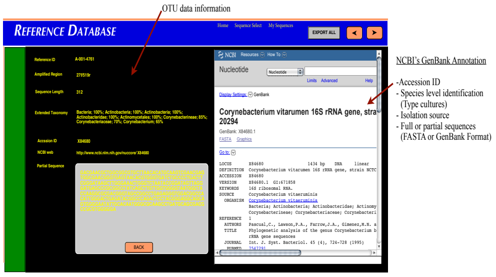
RFLP length (bp) 213 312 388 438 458

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.



OTU data information

Reference ID A-001-4761

Amplified Region 27f/519r

Sequence Length 312

Extended Taxonomy Bacteria; 100%; Actinobacteria; 100%; Actinobacteria; 100%; Actinobacteridae; 100%; Actinomycetales; 100%; Corynebacterineae; 85%; Corynebacteriaceae; 70%; Corynebacterium; 65%

Accssion ID X84680

NCBI web http://www.ncbi.nlm.nih.gov/nuccore/ X84680

Partial Sequence

```
GACGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAACGAG
TGGCGAACGGGTGAGTAACACGTGGGTGATCTGCCCCTGACTT
CGGGATAAGCTTGGGAAACTGGGTCTAATACCGGATAGTTAGG
GATGAGCCCGCGGCCTATCAGCTTGTTGGTGGGGTAATGGCTA
CCAAGGCGCGACGGGTAGCCGGCCTGAGAGGGTGGACGGCCA
CATTGGGACTGAGATACGGCCCAGACTCCTACGGGAGGCAGCA
GTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCGACG
CCGCGTGGGGGA
```

Corynebacterium vitarumen 16S rRNA gene, stra
20294

GenBank: X84680.1

```
LOCUS       X84680                1434 bp    DNA     linear
DEFINITION  Corynebacterium vitarumen 16S rRNA gene, strain NCTC
ACCESSION   X84680
VERSION     X84680.1  GI:671858
KEYWORDS    16S ribosomal RNA.
SOURCE      Corynebacterium vitaeruminis
  ORGANISM  Corynebacterium vitaeruminis
            Bacteria; Actinobacteria; Actinobacteridae; Actinomy
            Corynebacterineae; Corynebacteriaceae; Corynebacteri
REFERENCE   1
  AUTHORS   Pascual,C., Lawson,P.A., Farrow,J.A., Gimenez,M.N. a
  TITLE     Phylogenetic analysis of the genus Corynebacterium b
            rRNA gene sequences
  JOURNAL   Int. J. Syst. Bacteriol. 45 (4), 724-728 (1995)
   PUBMED   7547291
```

NCBI's GenBank Annotation

-Accession ID
- Species level identification
  (Type cultures)
- Isolation source
- Full or partial sequences
  (FASTA or GenBank Format)

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.

ACKNOWLEDGEMENT

The author would like to extend her deepest gratitude to her supervisor, Assoc. Prof. Hiroyasu Satoh for his continual guidance, encouragement and immense support given throughout the study period. Beside, equal appreciation goes to Prof. Takashi Mino for his constructive comments and valuable advices given during this study.

The author is equally grateful to her co-supervisor, Dr. Yukio Koibuchi for his valuable advices and assistance extend throughout the study. And, equal appreciation goes to Prof. Masahira Hattori, Dr. Kenshiro Oshima and their team at Centre for Omics and Bioinformatics, Graduate School of Frontier Sciences for providing their technical support for the study.

At the same time, author would like to express her gratitude to Dr. Mamoru Oshiki and all the laboratory members who helped for providing the initial materials and technical support at the very beginning of the study.

Sincere appreciation is expressed to Asian Development Bank (ADB-JSP) for granting her the prestigious scholarship to pursue her study in The University of Tokyo.

Words of appreciation also extended to her tutor, Mr. Masanori Takanashi for his support for making her stay in Japan much easier and memorable. At the same time, author expresses her gratitude Mrs. Yuriko Takeuchi, the secretary for her support and care express during the period of study.

In addition, sincere thanks goes to all my friend in the Mino-Satoh Laboratory for their encouragements, support and author wish them every success in their lives.

Finally, heartfelt gratitude to her beloved parents for their well wishes and her two brothers giving her a strong support and guidance throughout her life.

LIST OF ACRONYMS

bp	base pair
BLAST	Basic Local Alignment Search Tool
DDBJ	DNA Data Bank of Japan
DGGE	Denaturing Gradient Gel Electrophoresis
DNA	Deoxyribo Nucleic Acid
EMBL	European Molecular Biology Laboratory
FISH	Fluorescence In Situ Hybridization
NCBI	National Center for Biotechnology Information
NJ	Neighbor Joining
OUT	Operational Taxonomic Unit
OTUMAMi	Operational Taxonomic Unit Management And Mining
PCR	Polymerase Chain Reaction
QIIME	Quantitative Insight Into Microbial Ecology
RDP	Ribosomal Database Project
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribo Nucleic Acid
rRNA	ribosomal Ribo Nucleic Acid
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SBR	Sequence Batch Reactor
T-RFLP	Terminal Restriction Fragment Length Polymorphism
TRFs	Terminal Restriction Fragments
WWTP	Wastewater Treatment Plant

LIST OF FIGURES

LIST OF TABLES

Table of Contents

Chapter 1

Introduction

1.1 Background

Water environment is the backbone for beginning of life and human civilization. In history human civilization was initiated by communities who gathered around areas with water resources. Gradually their population increased and formed cities. The dramatic increase of population caused not a few environmental problems. For example, disposal of solid and liquid wastes caused water pollution, scarcity of safe water, and threats to public health. As a solution, simple water treatment practices were introduced to highly populated areas. With the improvement of those practices, sewer pipelines and sewage treatment plants were installed in order to remove polluted water coming from households and industries in large scale and further to treat it. In Late 1890's to early 20th century, biological wastewater treatment processes such as trickling filter and activated sludge processes were developed, and today, different kinds of biological wastewater treatment processes are in use.

In definition "wastewater" is the used water. Domestic wastewater contains organic matter, nutrients like nitrogen (N), Phosphorus (P), toxic compounds and lower level of dissolved oxygen. The goal of wastewater treatment is often the removal of these pollutants. Activated sludge process is one of the well developed biological wastewater treatment processes for this purpose.

The activated sludge process was developed by engineers in Great Britain in 1913. Experiments on treating sewage in a draw-and-fill reactor (the precursor to today's sequencing batch reactor) effectively removed organic pollutants in sewage. Researchers thought that the sludge had been activated, and they named the process "activated sludge process". The process structure was well designed to remove organic pollutants. And the main players in pollutant removal were soon found to be different microorganisms in activated sludge.

Numerous studies have been done to clarify microbial communities in activated sludge, as significant part of pollutants is removed by the activity of microorganisms.

Until 1980s, microbial identification methods depended significantly on isolation followed by physiological analyses and morphological observation of microorganisms. In late 1980s to early 1990s, molecular microbiology, which is based on genetic information of microorganisms, was developed and was introduced to the analysis of microbial communities in different environments.

Molecular methods were also introduced to analyse microbial communities in activated sludge. Molecular methods that are often applied for the activate sludge samples are Terminal Restriction Fragment length Polymorphism (T-RFLP), Denaturing Gel Gradient Electrophoresis (DGGE) and Florescence in-situ Hybridization (FISH).

The data analysis and interpretation is the least developed factor of those methodologies. Thus, present study aims at novel high-throughput DNA sequencing approach; "Pyrosequencing" for microbial community identification. Difficulties in sequence data handing and interpretation concerned as a challenge and tried a computational approach as solution. Study mainly concerns on preparation of samples for sequencing studies using activated sludge samples collected from different wastewater sources. Then, perform sequence data analysis and interpretations by computational approach. Rapid identification and quantification of microbial community structures is one challenging topic in sequence analysis studies, which will be greatly address in present research work.

At present, scientists struggle to track vast communities of bacteria found different environment and characterize by molecular techniques like DNA probing, Cloning, Polymerase chain reactions (PCR) and DNA sequencing. New methods widely used and today we have several DNA sequence databases (GenBank, EMBL, DDBJ) for identified bacterial strains. Similar approaches were also applied to activated sludge sources for better understanding of microbial population, which were previously left unnoticed with conventional methods (Eikelboom, 1975).

1.2 Objectives and Scope

The background of present study is application of 16S rRNA pyrosequencing approach for activated sludge systems. The key challenges in pyrosequencing are the management of massive DNA sequence data sets and efficient data interpretation in graphical manner. The development computational tools were initiated as a solution for these challenges.

It is for combination of sequence data management and interpretation. Increased efficiency for data analysis regardless the number of reads is another aim of present study and finally targeting on the development of a robust computational workflow which can express the power of 16S pyrosequencing approach for microbial community analysis.

The objectives of present study,

1. Development of computational workflow for 16S rRNA pyrosequencing
    - Data analysis and interpretation methodology for management of massive sets of 16S rDNA pyrosequences obtained form activated sludge samples

2. Development of floc library
    - Data organization tool for studying morphological and microbial data for flocs in activated sludge

3. Development of reference DNA sequence database
    - Reference sequence selection and sequence data repository for activated sludge microorganisms

1.3 Composition of Thesis

The composition of thesis,
Chapter 1 – Background of present study and complete framework explaining major and minor experimental procedures.

Chapter 2 – Literature review on principles and microbiological importance of biological wastewater treatment process, microbial community identification methodologies for activated sludge using conventional and molecular techniques and 16S rRNA pyrosequencing methodologies.

Chapter 3 – Experimental procedures for sample preparation and arrangement of sequence data for analysis

Chapter 4 – Explicit description on sequence data analyses based on different case studies and development of workflow for pyrosequencing data management.

Chapter 5 – Experimental procedures for activated sludge floc isolation, microbial community analysis by RFLP and development of floc library

Chapter 6 – Explanation sequence data handling for the development of reference sequence database.

Chapter 7 – Summary on complete work performed to in line with objectives. And, recommendations and improvements for future studies.

1.4 Research Framework

Application of 16S rRNA pyrosequencing for activated sludge sources is the basis oof the present study. Sludge samples collected at different wastewater treatment plants, industrial wastes and laboratory scale reactors were used for preparation of pyrosequence test sequences.

16S rDNA Pyrosequencing Workflow: Obtained sequences were used as the initial material for the development of workflow for sequence data management and interpretation. This has three main steps, *Experimental procedures, trial and errors* for data analysis and development of *computational approach.*

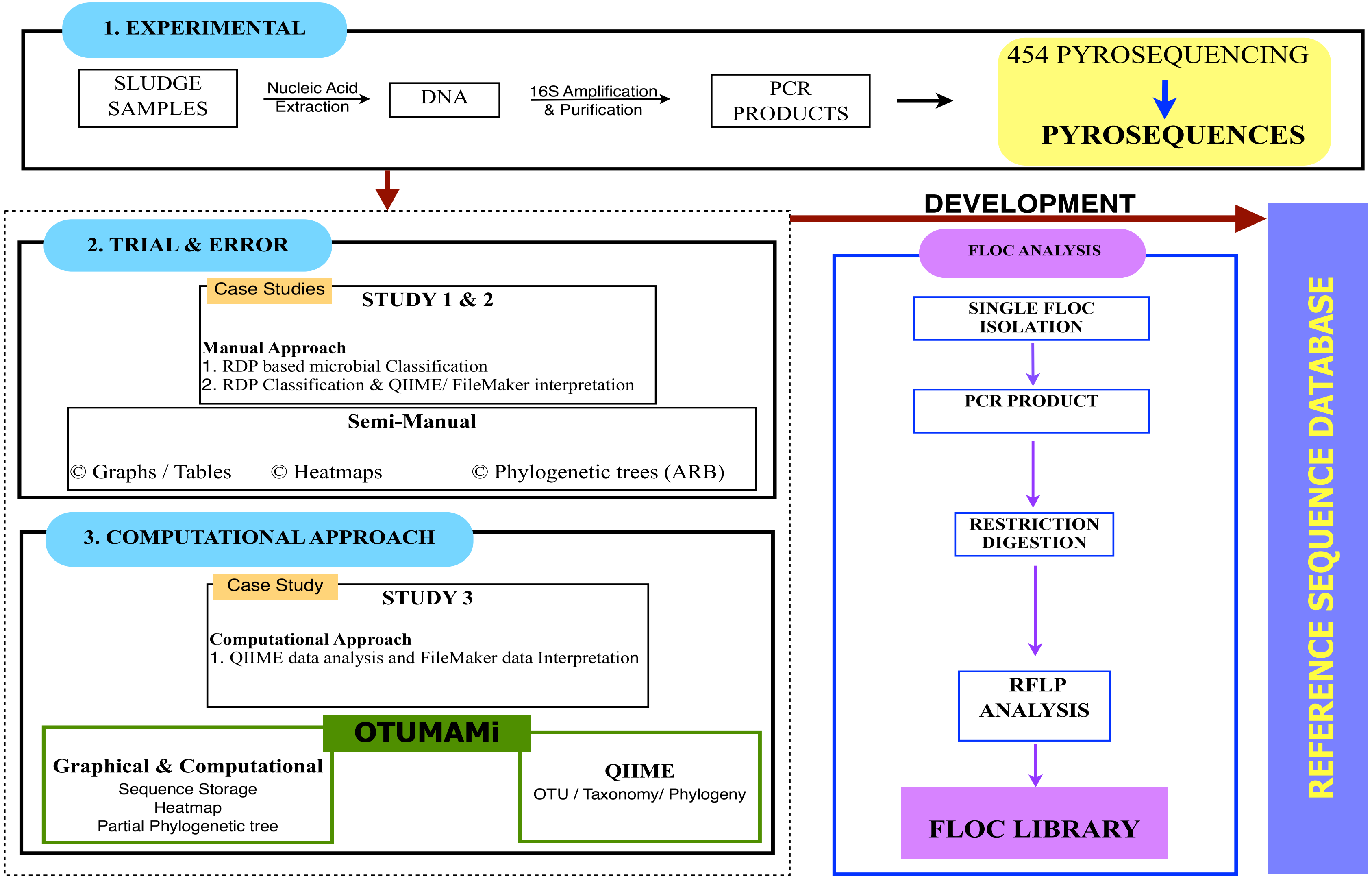
Computational approach (OTUMAMi) is for providing a single platform for,

- **Pyrosequence data analysis** – Quality assessments, Operational Taxonomic Units (OTUs) identification, Taxonomic assignments and phylogenetic tree construction
- **Sequence data interpretation** – Quantitative compositional analyses, identification of representative OTUs, graphical data interpretation (Charts, Heatmaps and phylogenetic trees).

Floc Library: Activated sludge flocs with different shapes, size were isolated and their microbial communities were determined by RFLP.

Reference Database: Identified major microbial community structures were further classified up to Genus level taxonomic assignment by Classifier tool in Ribosomal Database Project (RDP). Well-resolved taxonomic affinities with corresponding partial sequence data were stored in a simple database format.

In summary, present study was successful in exhibiting the necessity of workflow for handling of massive 16S pyrosequences in a computational manner. The role of computational tool was confirmed and major microbial communities were identified at a higher resolution. Their sequence data stored in reference sequence database for the development of database resource only for activated sludge microorganisms in future.


1. EXPERIMENTAL
SLUDGE SAMPLES
Nucleic Acid Extraction
DNA
16S Amplification & Purification
PCR PRODUCTS
454 PYROSEQUENCING
PYROSEQUENCES
2. TRIAL & ERROR
Case Studies
STUDY 1 & 2
Manual Approach
1. RDP based microbial Classification
2. RDP Classification & QIIME/ FileMaker interpretation
Semi-Manual
© Graphs / Tables
© Heatmaps
© Phylogenetic trees (ARB)
3. COMPUTATIONAL APPROACH
Case Study
STUDY 3
Computational Approach
1. QIIME data analysis and FileMaker data Interpretation
OTUMAMi
Graphical & Computational
Sequence Storage
Heatmap
Partial Phylogenetic tree
QIIME
OTU / Taxonomy/ Phylogeny
DEVELOPMENT
FLOC ANALYSIS
SINGLE FLOC ISOLATION
PCR PRODUCT
RESTRICTION DIGESTION
RFLP ANALYSIS
FLOC LIBRARY
REFERENCE SEQUENCE DATABASE

Chapter 2

Review of Literature

2.1 Biological Wastewater Treatment Process

Wastewater treatment is one of the most important biotechnological processes, which is used worldwide to treat municipal and industrial sewage. The biological component of process is the consortium of microorganisms mainly dominated by variety bacterial communities. Bacteria consume organic compounds for their supply of carbon and energy while helping the degradation of complex organic components flows with domestic and industrial wastewater sources (Langer, 1969).

Sewage from households, hospital and industries are collected into large volume tanks and treated at wastewater treatment plants. Method of treatment basically depends on type of sewage (Industrial or municipal) and its performances also vary from plant to plant. As a result, many kind of wastewater treatment methodologies were developed, ex. Trickling filter methods, Conventional activated sludge process, and Membrane filter methods.

The present study focused on activated sludge process, one of biological wastewater treatment methods with a significant ecological and economical impact (Mino, 2000).

The Primary target is,

- Generating a biological floc that is easy to settle

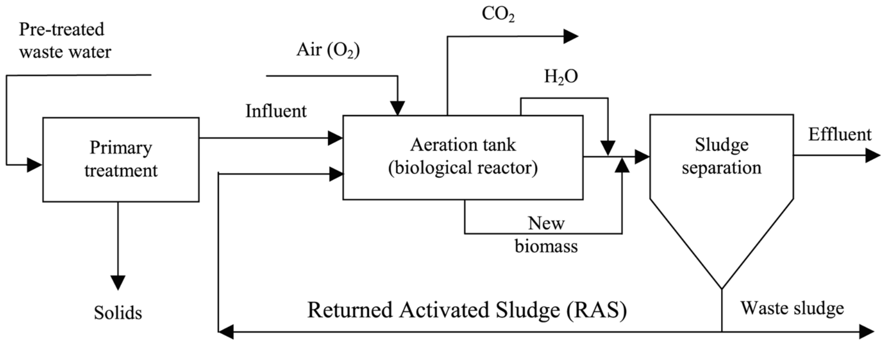
Biological floc attached with microorganisms, organic matter and extra cellular substances. This is main method for removing organic matter in sewage.

In addition,

- Oxidizing nitrogenous matter and removing phosphate

Mainly ammonium and nitrogen in biological materials oxidized as nitrogenous matter. The nutrient removal efficiency depends on the operational conditions.

Due to above reasons understanding microbial ecology in activated sludge became interesting and researchers became more keen on microbial identification and characterizing them for optimization of treatments processes.



Source: Adapted from Kiely (1997)

Fig.2.1 Layout of activated sludge system showing key processes and components

According to Fig. 2.1,

- Primary treatment – removing solid waste by settling
- Aeration tank - Injecting air (or oxygen) in the mixed liquor
  - Mixed liquor – wastewater plus mass of biological content
- Sludge separation - also referred to as "Final clarifier" or "secondary settling tank"
  - Allow the biological flocs to settle and separates biological sludge from the clear treated water (Effluent)

And, treatment of nitrogenous matter or phosphate involves additional steps where the mixed liquor is left in anoxic condition where is no residual dissolved oxygen (Beychok and Milton, 1967). Therefore, many forms of activated sludge systems were designed for applying additional treatments and enhancing the quality of treatment. Oxidation Ditch, EBPR (Enhanced Biological Phosphorus Removal), A2O (Anaerobic Aerobic and Oxic) and AO(Anoxic Oxic) are few of the improved processes.

2.2 Microbiology of Wastewater Treatment

In all plant types, prokaryotic microorganisms dominate and are responsible for the nutrient conversions. On the other hand, certain microorganisms cause the most frequently encountered problems in wastewater treatment like activated sludge bulking and foaming. Consequently, the efficiency and robustness of a waste mainly depend on the composition and activity of its microbial community (Eikelboom, 1975).

2.2.1 *Microorganisms and their functions*

Typically, most part of microorganisms in activated sludge is bacteria. The main role of microorganisms in activated sludge are (1) removal of organic matters, (2) removal of nutrients, and (3) formation of well-settlable flocs.

(1). **Removal of organic matter**: Bacteria playing key role activated sludge process by consuming biodegradable material such as proteins, carbohydrates, fats and many other compounds. Enzymes help bacteria in the process of breaking down nutrients, and in rebuilding broken down nutrients into the new compounds that they require for growth and reproduction.

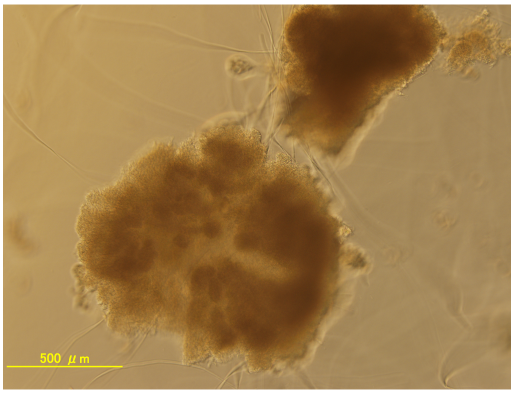
(2). **Removal of nutrient and Nutrient conversions**: The mechanisms of nutrient removal by microorganisms could be divided into two: removal by assimilation and others. Microorganisms require certain nutrients for growth. The basic nutrients of abundance in normal raw sewage are carbon (C), nitrogen (N), phosphorus (P), with the ratio of C:N:P ratio approximately equal to 100:10:1. Certain form of bacterial communities present in sludge responsible in nutrient conversions: Nitrification/ de-nitrification and phosphorus removal.

(3). **Formation of biological flocs**: As bacteria begin growing, they generally develop into small chains or clumps. Since they are very active and motile, it is difficult for them to settle. Also, due to mixing the small chains or clumps are broken up and the bugs are dispersed, and they will not flocculate or settle.

As the sludge is allowed to age, the microorganisms loose their motility and accumulate more slime (EPS- Extracellular Polymeric Substances). Then the clumps and chains are better able to stick together and grow bigger and bigger until they form a floc. If the organisms are allowed to develop properly, under the right conditions, the floc gets large and compact and begins to settle.

Activated sludge floc formation and structure

Floc formation takes place as a result of the active bacteria precipitating out EPS, which then bind the various components together. In a good sludge floc, the filamentous bacteria and other components act as the "backbone" that holds the sludge floc together.



In some cases, the presence of filamentous bacteria - whether they float freely between the sludge flocs or project between them - may prevent further flocculation from taking place, and the sedimentation properties from being drastically impaired. The sedimentation properties of the sludge depend to a high degree on the ability of the bacteria to "clump together" (form flocs). Bacteria that have this ability are known as floc-forming bacteria and effective floc formation is the basis for satisfactory separation during the sedimentation phase.

Therefore, structure of the sludge floc and the concentration of filamentous bacteria in the active sludge can vary considerably from plant to plant, and reflect differences in the plant types, operation and wastewater composition. This often leads to variations in the sedimentation properties of the active sludge.

Microorganisms with different functions in activated sludge

1) PAOs

Under Aerobic condition,

Phosphate → Poly-P ***…(Poly-P Accumulation)***

Polyphosphate Accumulating Organisms (PAOs) consume stored organic matter for growth (biomass production) and regeneration of energy (ATP). With the produced energy they take up the phosphate and regenerate Poly-P and store it.

Under Anaerobic condition,

Poly-P → Phosphate ***...(Phosphate Production)***

PAOs utilize Poly-Phosphate as the energy source and take up the organic matter

2) GAOs

Glycogen accumulating organisms (GAOs) are another type of bacterial community found in activated sludge systems. Glycogen acts as their main sources of energy under anaerobic condition. This will indirectly effect on the PAOs functions at the same condition since they too oxidize the organic matter. This cause a competition between and GAOs and PAOs which finally leading to poor waste treatment.

3) Nitrifiers

NH4+ → NO2- → NO3- ***….(Nitrification)***

Nitrification has two steps;

Ammonia oxidizers convert Ammonia to nitrite and nitrite oxidizers consume nitrite and produce nitrate. Phylogenetically, Ammonia oxidizers and nitrite oxidizers are totally different.

Nitrate (NO3-) → N2 ***…(De-nitrification)***

De-nitrifying bacteria convert nitrate and nitrite to dinitrogen gas while consuming organic matters. By combining nitrification and de-nitrification reactions, nitrogen in wastewater is converted to dinitrogen gas, and nitrogen removal is achieved.

4) Filamentous microorganisms and bulking

Dense growth of filamentous bacteria communities cause sludge bulking that leads to failures in wastewater treatment process. Thus maintenance of stable community seems to be essential. However, this is the most challenge wastewater engineering facing in maintain the optimum performances at full scale wastewater treatment.

2.2.2 *Methods for Microbial Community Analysis*

Although biological wastewater treatment has been used for more than a century, research on the microbiology of this process suffered from severe methodological limitations (Wagner et al., 1993). Understanding of bacterial community structure became more pronounced after introduction of different conventional and molecular techniques in wastewater microbiology and to identify the microbial key players for the different process types.

(a). Conventional Techniques

The culture-dependent techniques, in which microorganisms are isolated by culturing on plate or in tube, were successfully applied to isolate and identify pathogenic and health-related bacteria. They were also applied to many environmental samples including activated sludge samples. Those were laborious methods where morphological characters were identified by microscopic views.

> ***Microbial Plate Culture****: A microbiological culture method of multiplying microbial organisms by letting them reproduce in predetermined culture media under controlled laboratory conditions. Microbial cultures are used to determine the type of organism, its abundance in the sample. The characteristics were identified using microscopic views and classify accordingly.*

Microbial cultures are foundational and basic diagnostic methods used extensively for pure culture preparations and community characterization. The development of bacterial systematics were started and publication of resources like “The Bergey’s Manual of Systematic Bacteriology (1923)” which is the main resource for determining the identity of bacteria species, utilizing every characterizing aspect, also based on those conventional techniques. The manual itself explains classification of bacteria based on their structural and functional attributes by arranging them into specific familial orders.

However, low percentage of activated sludge bacteria culturable on general plating media and some tend to grow on selective nutrient media. This one major limitation in conventional techniques and showed that detection and identification of less abundant and morphologically less characteristic bacteria require different techniques.

(b). Molecular Techniques

In the last decade a set of molecular tools has been developed and applied to investigate the microbial community structures and quantify their compositions. Presently, culture independent molecular methods have emerged as indispensable tools for studying microbial community structure and dynamics in natural habitats, since they allow a closer look at microbial diversity that is not reflected by culturing techniques. Microbial diversity in activated sludge also explored by novel culture independent tools targeting the genetic components (DNA/ RNA) of bacteria.

The advent of molecular biology in the 1980s contributed a set of powerful new tools that have helped microbiologists to detect the smallest variations within microbial species and even within individual strains. This has added an entirely new dimension to a science that was in danger of becoming constrained by its reliance on traditional laboratory techniques (Head et al., 1998).

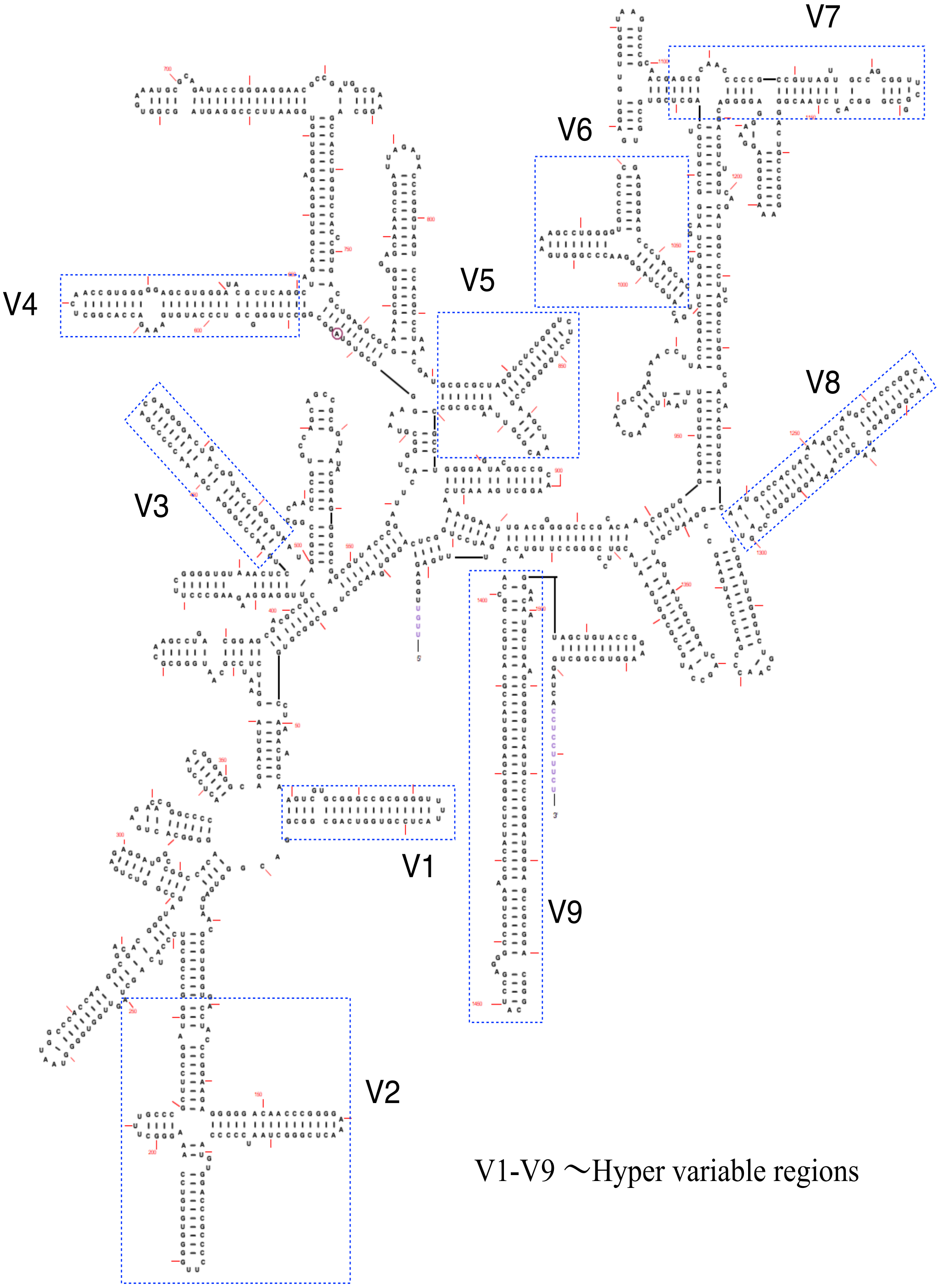
2.2.3 *Molecular techniques for microbial typing in activated sludge*

The improved DNA extraction and PCR amplification methods contribute more on the advanced development of molecular techniques we are using today. New methods are extensively using to characterize microbial populations involved in different activated sludge systems and those methods include clone library construction, Terminal Restriction Fragment Length Polymorphisms (T-RFLP), Denaturing Gradient Gel Electrophoresis (DGGE) and Fluorescent in situ Hybridization (FISH). At latest, full-cycle 16S rRNA sequencing approach also working hand in hand for comprehensive analysis of bacterial diversity.

Various genes are used in molecular systematic studies of microorganisms. Most widely used and useful for defining phylogenetic affinities are the genes encoding 16S rRNA in bacterial genome. The small sub unit of ribosomal RNA (SSU rRNA) genes have been used extensively for sequence based analysis due to following reasons:

- Universal distribution
- Functionally consistent
- Sufficiently conserved
- Adequate length (1542bp) for sequence analyses

Carl Woese at the University of Illinois pioneered the use of SSU rRNA for phylogenetic studies in early 1970s. This gave rise to the novel classification of living organisms with three domains, Bacteria, Archaea and Eukarya.



(Source: 16S rRNA secondary structure drawn by RNAD2MAP Programme from Comparative RNA Web Tool)

Fig.2.2 Ribosomal RNA Secondary structure from *Escherichia Coli* (Bacteria)

Microbial community analyses and phylogenetic studies using DNA sequences relies heavily on the polymerase chain reaction (PCR) to obtain sufficient copies for efficient sequencing and further analysis using FISH, RFLP and DGGE.

Specific oligonucleotide primers (Lane, 1991) are well designed for amplification of gene of interest and standard primers exist for many highly conserved regions from 27f to 1454r for 16S rRNA gene in bacteria.

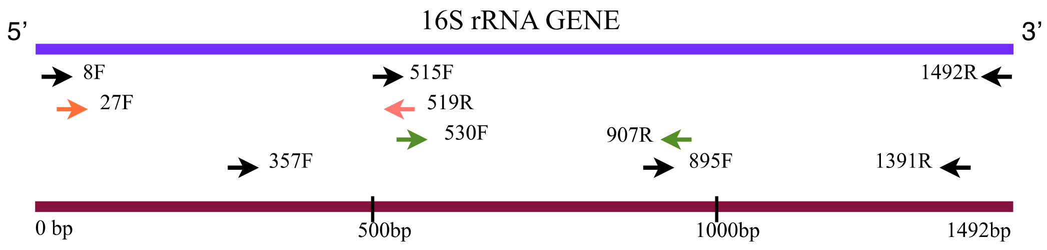


Fig.2.3 Primer map for the 16S SSU rRNA gene (Universal & specific primers)

Table 2.1 primer sequences for 16S rRNA gene amplification (*Numbered primers are named for the approximate position on *E.coli*)

Primer*	**Sequence(5'-3')**	**TargetGroup**	**Reference**
8F	AGAGTTTGATCCTGGCTCAG	Universal	Turner et al.1999
27F	AGAGTTTGATCMTGGCTCAG	Universal	Lane et al.1991
CC[F]	CCAGACTCCTACGGGAGGCAGC	Universal	Rudi et al.1997
357F	CTCCTACGGGAGGCAGCAG	Universal	Turner et al.1999
515F	GTGCCAGCMGCCGCGGTAA	Universal	Turner et al.1999
530F	GTGCCAGCAGCCGCGG	Universal	Weisburg et al.1991
1237F	GGGCTACACACGYGCWAC	Universal	Turner et al.1999
519R	GWATTACCGCGGCKGCTG	Universal	Turner et al.1999
CD[R]	CTTGTGCGGGCCCCCGTCAATTC	Universal	Rudi et al.1997
907R	CCGTCAATTCMTTTRAGTTT	Universal	Lane et al.1991
1391R	GACGGGCGGTGTGTRCA	Universal	Turner et al.1999
1492R(l)	GGTTACCTTGTTACGACTT	Universal	Turner et al.1999
1492R(s)	ACCTTGTTACGACTT	Universal	Lane et al.1991

A. Terminal Restriction Fragment Length Polymorphisms (T-RFLP)

Terminal Restriction Fragment Length Polymorphism is another popular methodology extensively used for activated sludge samples that also depends on 16S rDNA PCR amplification.

T-RFLP analysis is a technique used to study complex microbial communities based on variation in the 16S rRNA gene. It is a culture- independent, rapid, sensitive and reproducible method of assessing diversity of complex communities without the need for any genomic sequence information.

Applications:

T-RFLP for exploring microbial community structure and community dynamics in response to changes in different environmental parameters or to study and quantify bacterial populations in diverse environments such as soil, marine and activated sludge systems

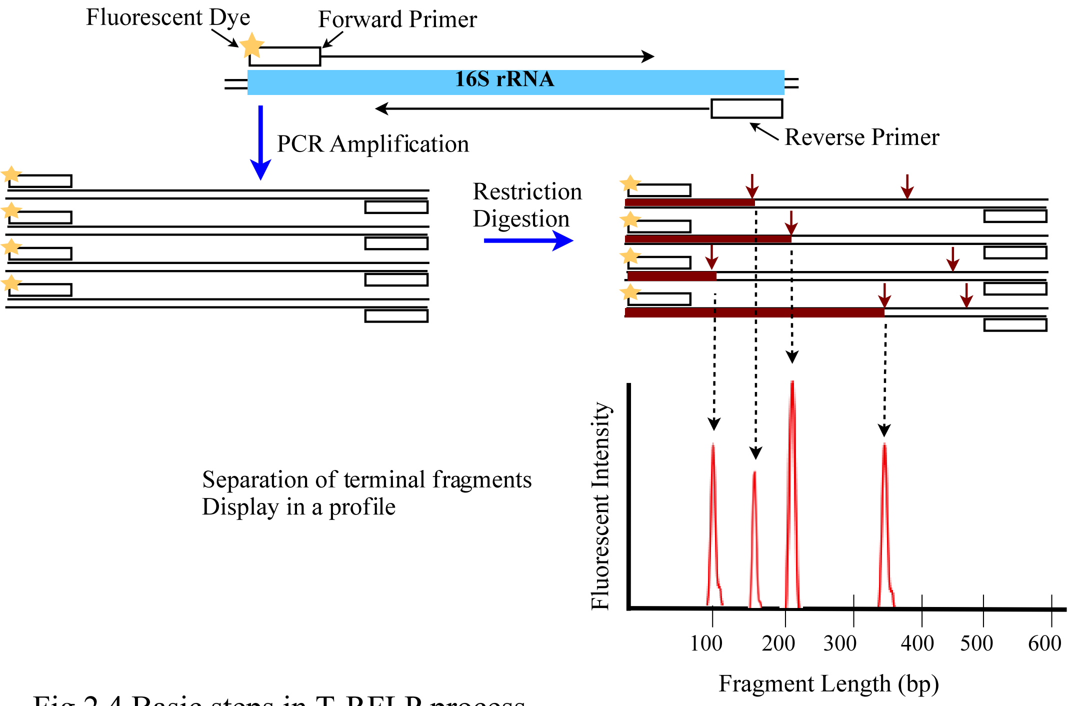


Fig.2.4 Basic steps in T-RFLP process.

A T-RFLP profile (Fig.2.5) plot against Terminal Restriction Fragment lengths (T-RFs) and different microbial communities were corresponded with T-RF length and abundance with peak intensity.

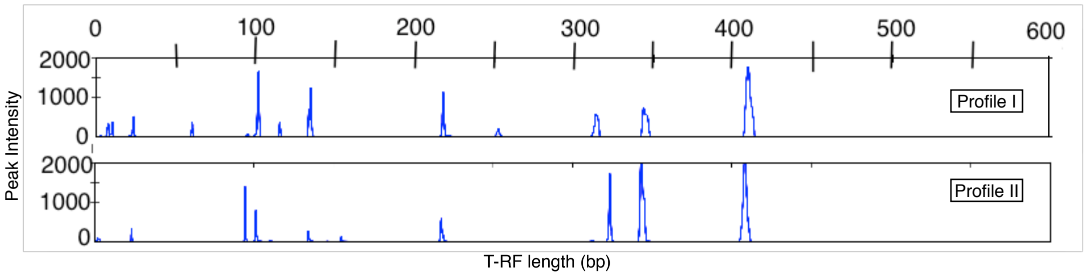


Fig.2.5 T-RFLP profiles showing peak pattern between peak intensity and T-RFs length (bp)

Studies outlines in Hiraishi et al., 2000; Blackwood et al., 2003; Dickie et al., 2007 and Slater et al., 2010 are done on exploring microbial ecology using T-RFLP in different environmental samples. Researchers were interested in developing data analysis methods (Marsh et al., 2000; Kent et al., 2003), confirming the reproducibility of T-RFLP profiles (Caffaro-filho et al, 2007) and application of different restriction enzymes or florescence dyes (Hiraishi et al., 2000; Pandey et al., 2007) for better understanding of microbial diversity. Therefore, T-RFLP became a powerful tool for microbial community identification and quantification. Later, found that it's ideal method for confirming microbial community dynamics in response to change of time, environment and condition.

Limitations:

Terminal Restriction Fragment Length Polymorphism (T-RFLP) analysis is one of the informative and widely used techniques for such studies. However, the method has a few limitations to predict microbial community structure with significant accuracy.

1. Variations found in observed and expected TRFs

Variations in real Terminal Restriction Fragment (TRF) length and expected TRF length is a one major limitation.. Therefore, similar TRFs can be owned by different community structures and finally let its peak intensities overlaps on each.

2. Poor data processing and analysis

And also, the least-well-defined technical aspect of T-RFLP is the data processing and analysis of profiles. A wide range of methods has been used emerged however still it's on the way to find an optimal procedure for comparing complex environmental T-RFLP profiles.

Then, comprehensive analyses on the bacterial diversities were further improved on so-called full-cycle 16S rRNA approach together with next generation sequencing methodologies.

2.3 Next Generation DNA Sequencing Approach

The field of DNA sequencing technology development has a rich and diverse history. However, the majority of DNA sequence production to date has relied on some version of the Sanger sequencing. It started early 1990s, DNA sequence production has almost exclusively been carried out with capillary-based semi-automated process. Sequencing is a ‘cycle sequencing’ reaction which cycles of template denaturation, primer annealing and primer extension are performed. Mixture of extension products is end labeled on the terminating ddNTP (terminal position). Sequence is determined by high-resolution electrophoresis (capillary- based polymer gel). Software translates these fragments into DNA sequence.

Over past ten years, the incentive for developing entirely new strategies for DNA sequencing has emerged with significant reductions in the cost of conventional DNA sequencing. Variety of molecular methods have been developed, whereby a broad range of biological phenomena can be assessed by high-throughput DNA sequencing (e.g., genetic variation, RNA expression, protein-DNA interactions and chromosome conformation). The 454 sequencing is one such promising high-throughput sequencing methodology commonly known as “454 Pyrosequencing” widely used for in metagenomic studies. (Shendure et al., 2008).

2.3.1 Development of Pyrosequencing

Pyrosequencing has been available to the scientific community since the mid-1990s as a genotyping tool. It is a method of DNA sequencing based on the "sequencing by synthesis" principle. It differs from Sanger sequencing, in that it relies on the detection of pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides (ddNTP). The technique was developed by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology in Stockholm, Sweden in 1996 (Ronaghi et al., 1996;1998; Nyren et al., 2007).

Currently, a limitation of the method is that the lengths of individual reads of DNA sequence are of 300-500 nucleotides which shorter than the 800-1000 obtainable with chain termination methods (e.g. Sanger sequencing). This can make the process of genome assembly more difficult, particularly for sequences containing a large amount of repetitive DNA. However, pyrosequencing is most commonly used for re-sequencing of genomes for which the sequence of a close relative is already available.

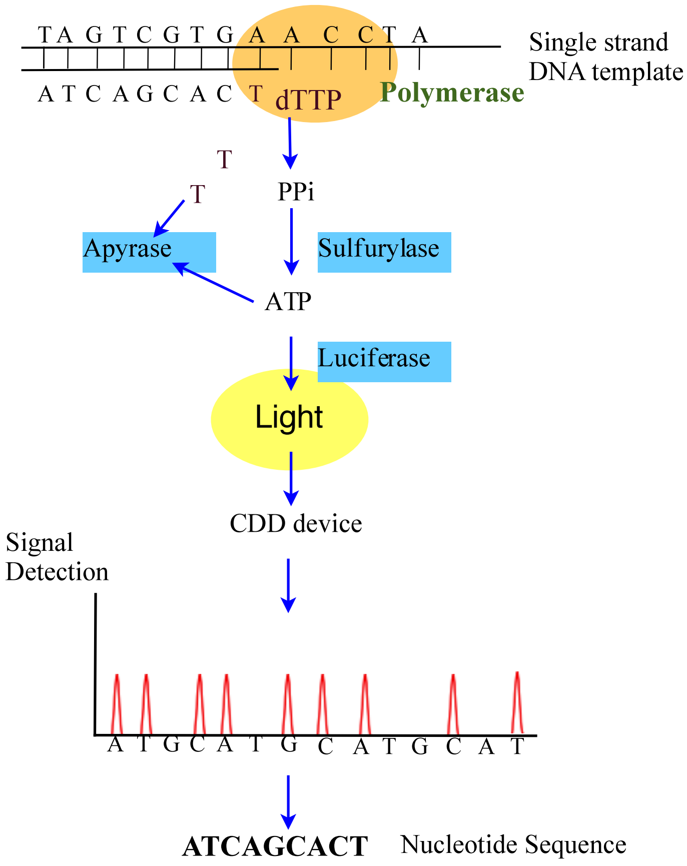


Fig.2.6 The mechanism of pyrosequencing process

2.3.2 Application of Pyrosequencing

Metagenomics

One main application of pyrosequencing that study on metagenomes which genetic material recovered directly from environmental samples.

Few studies reported with pyrosequencing approach are,

1. Use of barcoded primer and shared OTUs for identification of fecal bacterial communities in watersheds (Unno et al., 2010)
2. Microbial diversity profiling for human microbiome projects (Hamady and Knight, 2010)
3. Metagenomic analysis on EBPR sludge communities (Martin et al., 2006)
4. Pyrosequencing for identification of soil microbial diversity (Roesch et al., 2007)

Studies reported from Liu et al., 2007 gives a better understanding on pyrosequencing-based data analysis tools and methodologies. And, Shendure and Ji, 2008 gives a review on next generation sequencing protocols and it's applications and challenges.

Traditional microbiology and microbial genome sequencing rely upon cultivated clonal cultures. The use of pyrosequencing helped to bypass bacterial culturing and cloning enables studies of organisms that are not easily cultured in a laboratory as well as studies of organisms in their natural environment (Macro et al., 2010). The basic steps followed in metagenomic studies as follows,

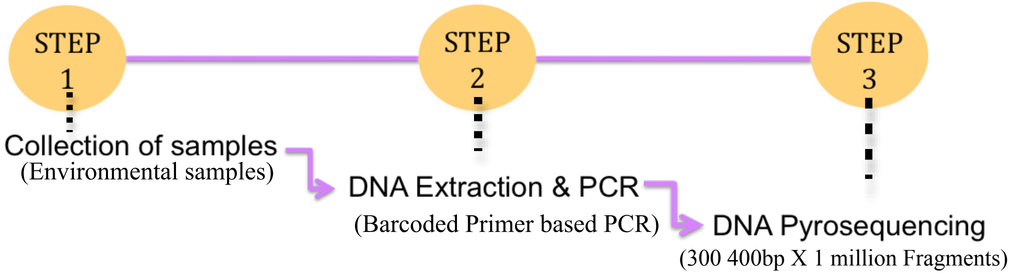


Fig.2.7 Preliminary steps for pyrosequencing aided metagenomic study.

The use of barcoded primers in PCR amplification is one unique feature in pyrosequencing protocol. This enables multiplexed sequencing; simultaneous analyze of multiple targets (samples) of DNA. Due to multiplexing sequencing cost being reduced. The generation of massive number of DNA fragments (~ 1 Million) enables the sequencing of both major and minor microbial genomes in samples of interest. Therefore high resolution taxonomic insight is provided along with Operational Taxonomic Units (OTUs).

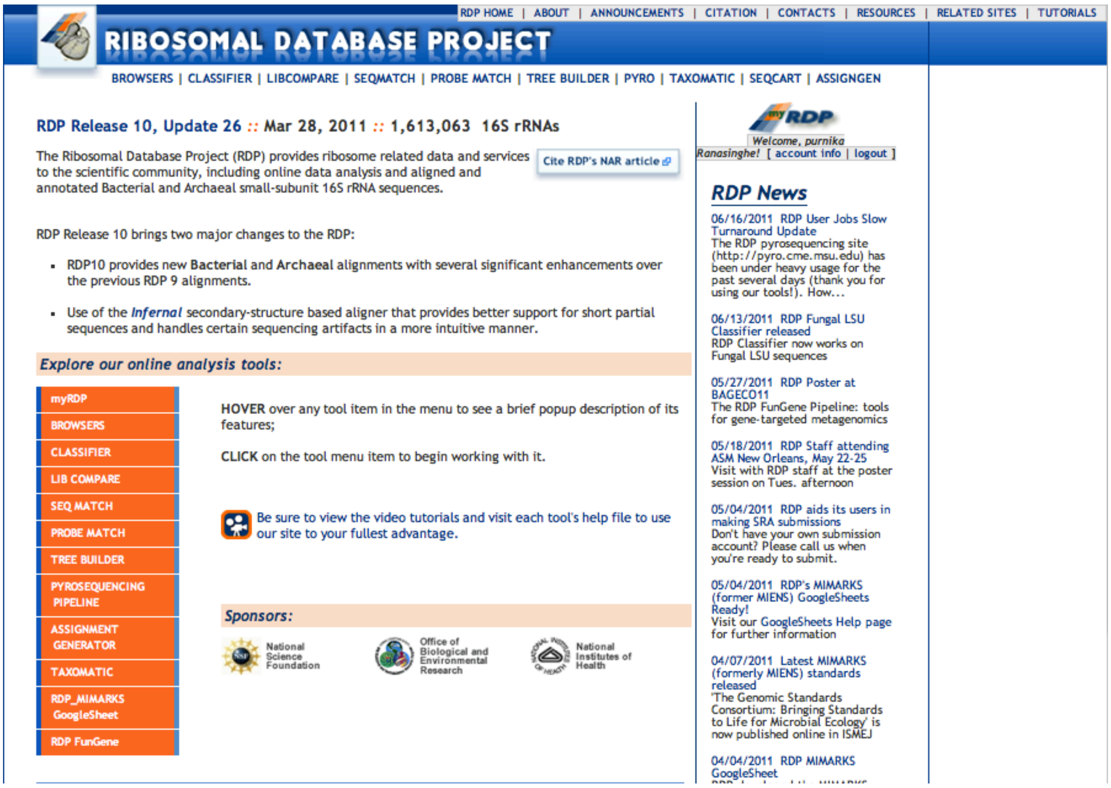
2.3.3 Data Analysis and Interpretation by Pyrosequencing

Next step of massive parallel Pyrosequencing is the data analysis and interpretation. Natively most of the molecular tools it-selves failed in providing a better data analysis or interpretation. Similarly pyrosequencing analysis also aided with several sequence analyses tools (RDP, Greengenes) and software (ARB-SILVA, WATERS and QIIME).

A. Ribosomal Database Project (RDP)

(Release 10)

The Ribosomal Database Project (RDP) provides researchers with quality-controlled bacterial and archaeal small subunit rRNA alignments and analysis tools. Pyrosequencing Pipeline is one such tool to support analysis of ultra high-throughput rRNA sequencing data. This pipeline offers a collection of tools that automate the data processing and simplify the computationally intensive analysis of large sequencing libraries. Details about RDP data and analytical functions can be found at http://rdp.cme.msu.edu/ (Cole et al., 2003; 2008, Petrosino et al., 2009)


RDP HOME | ABOUT | ANNOUNCEMENTS | CITATION | CONTACTS | RESOURCES | RELATED SITES | TUTORIALS
RIBOSOMAL DATABASE PROJECT
BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE BUILDER | PYRO | TAXOMATIC | SEQCART | ASSIGNGEN
Cite RDP's NAR article
myRDP
Welcome, purnika Ranasinghe! [account info | logout]
myRDP
BROWSERS
CLASSIFIER
LIB COMPARE
SEQ MATCH
PROBE MATCH
TREE BUILDER
PYROSEQUENCING PIPELINE
ASSIGNMENT GENERATOR
TAXOMATIC
RDP_MIMARKS GoogleSheet
RDP FunGene
National Science Foundation
Office of Biological and Environmental Research
National Institutes of Health

RDP Release 10, Update 26 :: Mar 28, 2011 :: 1,613,063 16S rRNAs

The Ribosomal Database Project (RDP) provides ribosome related data and services to the scientific community, including online data analysis and aligned and annotated Bacterial and Archaeal small-subunit 16S rRNA sequences.

RDP Release 10 brings two major changes to the RDP:

- RDP10 provides new **Bacterial** and **Archaeal** alignments with several significant enhancements over the previous RDP 9 alignments.
- Use of the *Infernal* secondary-structure based aligner that provides better support for short partial sequences and handles certain sequencing artifacts in a more intuitive manner.

Explore our online analysis tools:

HOVER over any tool item in the menu to see a brief popup description of its features;

CLICK on the tool menu item to begin working with it.

Be sure to view the video tutorials and visit each tool's help file to use our site to your fullest advantage.

Sponsors:

RDP News

06/16/2011 RDP User Jobs Slow Turnaround Update
The RDP pyrosequencing site (http://pyro.cme.msu.edu) has been under heavy usage for the past several days (thank you for using our tools!). How...

06/13/2011 RDP Fungal LSU Classifier released
RDP Classifier now works on Fungal LSU sequences

05/27/2011 RDP Poster at BAGECO11
The RDP FunGene Pipeline: tools for gene-targeted metagenomics

05/18/2011 RDP Staff attending ASM New Orleans, May 22-25
Visit with RDP staff at the poster session on Tues. afternoon

05/04/2011 RDP aids its users in making SRA submissions
Don't have your own submission account? Please call us when you're ready to submit.

05/04/2011 RDP's MIMARKS (former MIENS) GoogleSheets Ready!
Visit our GoogleSheets Help page for further information

04/07/2011 Latest MIMARKS (formerly MIENS) standards released
'The Genomic Standards Consortium: Bringing Standards to Life for Microbial Ecology' is now published online in ISMEJ

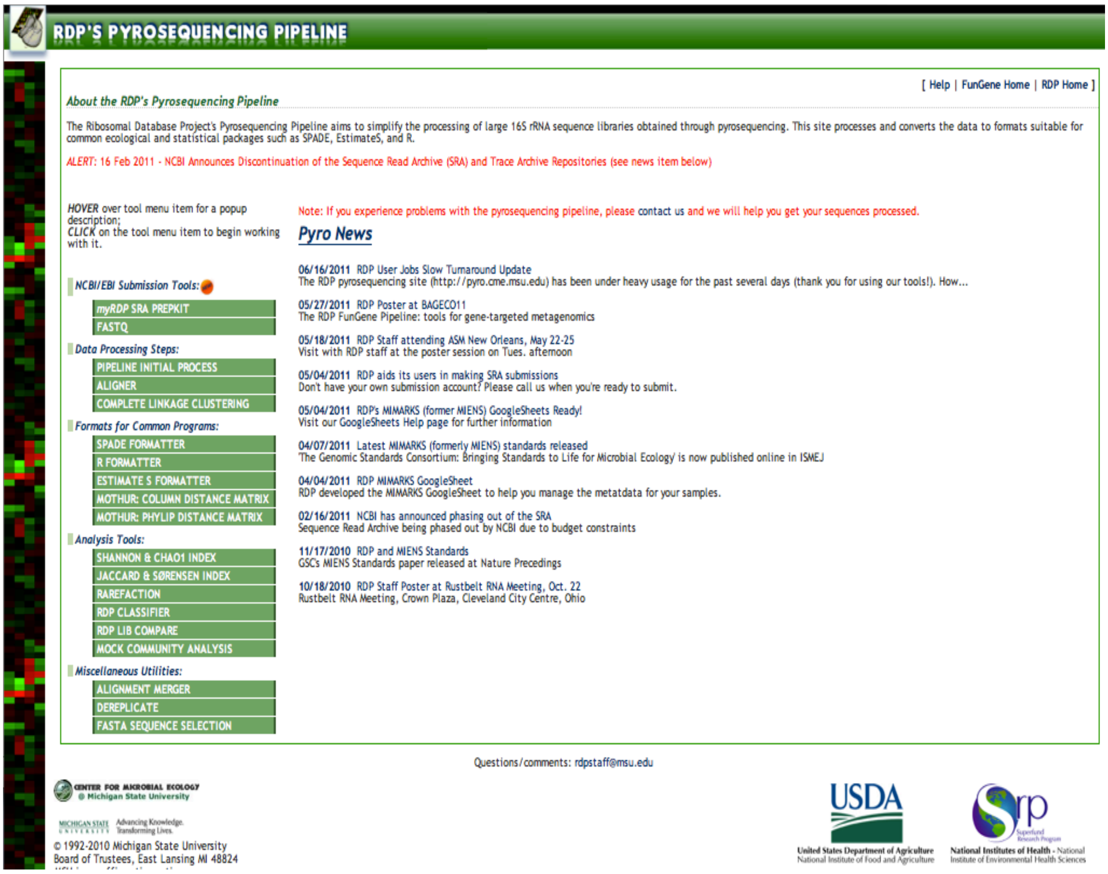
04/04/2011 RDP MIMARKS GoogleSheet

Fig.2.8 RDP homepage (source: http://rdp.cme.msu.edu)

The online ribosomal analysis application called "RDP's Pyrosequencing Pipeline" provides collection of tools to analyze With these new sequencing methodologies, computational analysis of the large numbers of sequences produced has become a major challenge. The new RDP Pyrosequencing Pipeline offers a collection of tools that automate the data processing and simplify the computationally intensive analysis of large sequencing libraries (Fig. 2.8).

Online data analysis (RDP's Pyrosequencing Pipeline)

- Sequence Alignment (RDP Aligner)
- Sequence Quality Analyses (Pipeline Initial Process)
- Sequence data storage (myRDP – personal sequence accounts)
- Annotation for Bacterial and Archaeal 16S rRNA sequences (RDP Classifier)
- Phylogenetic tree construction (Tree Builder)
- Selection of representative sequences (Dereplicate)
- Sequence Clustering (Complete Linkage clustering)
- Heatmap representation (Taxomatic Visualization tool)


RDP'S PYROSEQUENCING PIPELINE
[Help | FunGene Home | RDP Home]
NCBI/EBI Submission Tools:
myRDP SRA PREPKIT
FASTQ
Data Processing Steps:
PIPELINE INITIAL PROCESS
ALIGNER
COMPLETE LINKAGE CLUSTERING
Formats for Common Programs:
SPADE FORMATTER
R FORMATTER
ESTIMATE S FORMATTER
MOTHUR: COLUMN DISTANCE MATRIX
MOTHUR: PHYLIP DISTANCE MATRIX
Analysis Tools:
SHANNON & CHAO1 INDEX
JACCARD & SØRENSEN INDEX
RAREFACTION
RDP CLASSIFIER
RDP LIB COMPARE
MOCK COMMUNITY ANALYSIS
Miscellaneous Utilities:
ALIGNMENT MERGER
DEREPLICATE
FASTA SEQUENCE SELECTION
CENTER FOR MICROBIAL ECOLOGY @ Michigan State University
MICHIGAN STATE UNIVERSITY Advancing Knowledge. Transforming Lives.
USDA
United States Department of Agriculture National Institute of Food and Agriculture
Srp Superfund Research Program
National Institutes of Health - National Institute of Environmental Health Sciences

About the RDP's Pyrosequencing Pipeline

The Ribosomal Database Project's Pyrosequencing Pipeline aims to simplify the processing of large 16S rRNA sequence libraries obtained through pyrosequencing. This site processes and converts the data to formats suitable for common ecological and statistical packages such as SPADE, EstimateS, and R.

ALERT: 16 Feb 2011 - NCBI Announces Discontinuation of the Sequence Read Archive (SRA) and Trace Archive Repositories (see news item below)

HOVER over tool menu item for a popup description;
CLICK on the tool menu item to begin working with it.

Note: If you experience problems with the pyrosequencing pipeline, please contact us and we will help you get your sequences processed.

Pyro News

06/16/2011 RDP User Jobs Slow Turnaround Update
The RDP pyrosequencing site (http://pyro.cme.msu.edu) has been under heavy usage for the past several days (thank you for using our tools!). How...

05/27/2011 RDP Poster at BAGECO11
The RDP FunGene Pipeline: tools for gene-targeted metagenomics

05/18/2011 RDP Staff attending ASM New Orleans, May 22-25
Visit with RDP staff at the poster session on Tues. afternoon

05/04/2011 RDP aids its users in making SRA submissions
Don't have your own submission account? Please call us when you're ready to submit.

05/04/2011 RDP's MIMARKS (former MIENS) GoogleSheets Ready!
Visit our GoogleSheets Help page for further information

04/07/2011 Latest MIMARKS (formerly MIENS) standards released
'The Genomic Standards Consortium: Bringing Standards to Life for Microbial Ecology' is now published online in ISMEJ

04/04/2011 RDP MIMARKS GoogleSheet
RDP developed the MIMARKS GoogleSheet to help you manage the metatdata for your samples.

02/16/2011 NCBI has announced phasing out of the SRA
Sequence Read Archive being phased out by NCBI due to budget constraints

11/17/2010 RDP and MIENS Standards
GSC's MIENS Standards paper released at Nature Precedings

10/18/2010 RDP Staff Poster at Rustbelt RNA Meeting, Oct. 22
Rustbelt RNA Meeting, Crown Plaza, Cleveland City Centre, Ohio

Questions/comments: rdpstaff@msu.edu

© 1992-2010 Michigan State University
Board of Trustees, East Lansing MI 48824

Fig. 2.9 RDP's Pyrosequencing. Links for "Analysis Tools" and "Help" are given.

RDP itself has some limitations in processing large number of sequences for process. Since it's a online analyses tool shared at a centralized network, it usually has stringent limitations like CPU timing, Memory size and network bandwidth. In Pipeline processing model, set of applications are connected to each other such that output from one application becomes the input to one or more applications in the subsequent stage.

Therefore, bioinformaticians think that pipeline model is not an ideal design for analyzing sequence data. In addition, the process of file upload, analysis and download which may repeated at different stages is time consuming since the user needs to wait for output and must again upload the result again for the next stage of analysis. Another main limiting factor is that the pipeline tools cannot be customized and enhanced for any specialized data analysis since its running on another server.

Therefore, scientists became interested in Softwares that can download run on their own computers than uploading their data into another server.

B. Quantitative Insight Into Microbial Ecology (QIIME)

QIIME is a pipeline for performing microbial community analysis that integrates many analysis tools. QIIME can run on a laptop, a supercomputer, and systems in between such as multi-core desktops. It designed to investigate microbial diversity within and between samples using SSU (16S and 18S) rRNA gene sequences where gene sequences are generated by PCR sequenced with 454 pyrosequencing. Software can download from http://qiime.sourceforge.net/ and install according to instructions.

Basic approaches of microbial diversity analysis embedded in QIIME are,

- Sequence sorting and barcode identification
- Operational Taxonomic Units based clustering (OTUs)
- Representative sequence picking for OTUs
- OTU alignment by PyNast
- Taxonomy assignment by RDP classifier
- Phylogenetic tree building (Newick format)

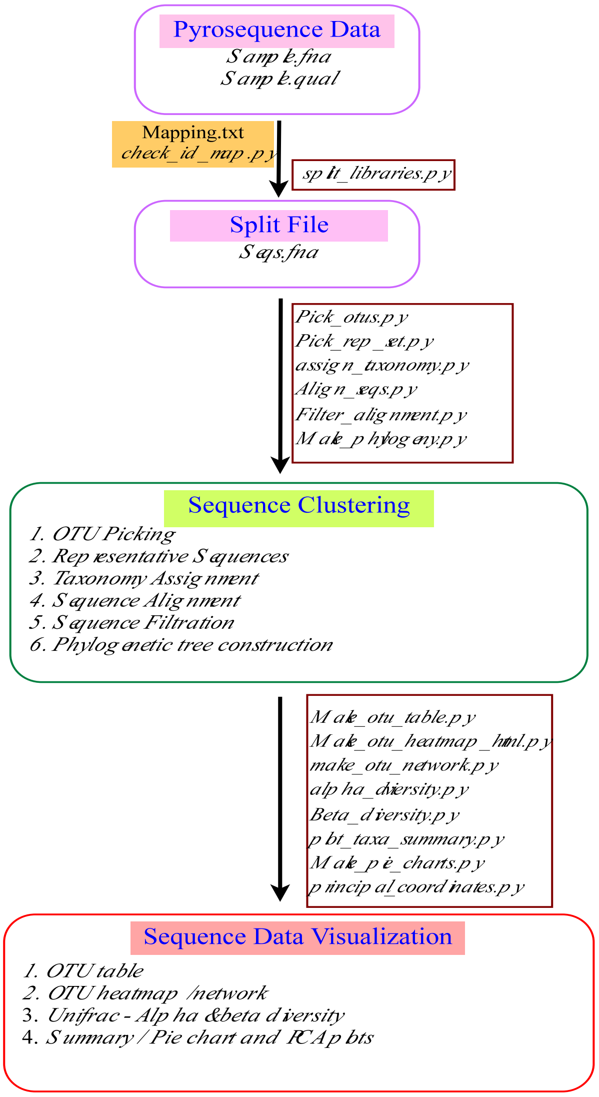
And, data visualization by,

- Unifrac diversity measurements (Alpha and beta diversity)
- OTU tables and heatmap
- OTU network/ Pie charts / Histograms and PCA plots

As shown in Fig.2.10, QIIME provides single platform for data analysis using commands allowing user to modify file downloading and saving options. Rather than re-implementing commonly used algorithms, QIIME can handle large datasets and perform quality check and chimers checking for choosing OTUs. It also allows faster integration of data that reduces the time spend on analysis.

QIIME allows user to use their own computational resources without uploading data into another server. However, it also runs on command line and that costs significant amount of time for learning, operating and performing functions and commands.

The present section expressed how next-generation DNA sequencing platforms are utilized in biological researches, how they work, their relative strengths and limitations together with emerging applications. Applications of sequence analysis tools and softwares are further explained in case studies conducted and more details find in Chapter 4 and 5.



Qiime command:

```
filter_alignment.py -i all_15.txt_rep_set_aligned.fasta -m lanemask_in_1s_and_0s -o filtered_alignment/
```

Filter_alignment.py : Python command
-i : Input File
-m : Mapping File
-o : Output File

Fig. 2.10 QIIME analysis process from sequence quality analysis to data visualization

2.4 Application of Relation Databases

A “Database” simply known as vast collection of information that require a computer and technical knowledge to access and mange data. And, the information in database is organized then one can find what they are looking for quickly and easily.

A relational database structures were built to matches data by using common characteristics found within the data set. The resulting groups of data are organized and are much easier for many people to understand. Relational databases are popular source of data storage specially for administration purposes, in banking, stock handling and even in small phone book applications. At present, relational database management protocols were applied for biological information such DNA sequences and protein structure data.

Following example helps to understand functionality of a relational database.

Table 01: A data set containing Operational taxonomic units (OTU) numbers, their partial sequence and sequence length.

Table 02: origin of sample, OTU numbers and their taxonomy.

Making relationships among tables is the main feature of building relational databases. The Fig.2.11 shows how the two tables connected to each other using OTU number as their common field.

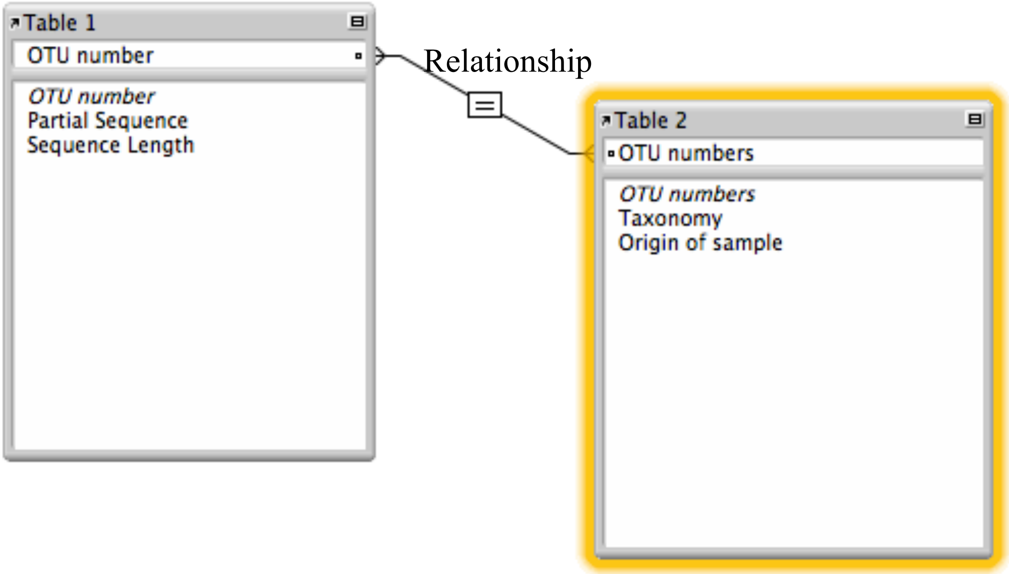
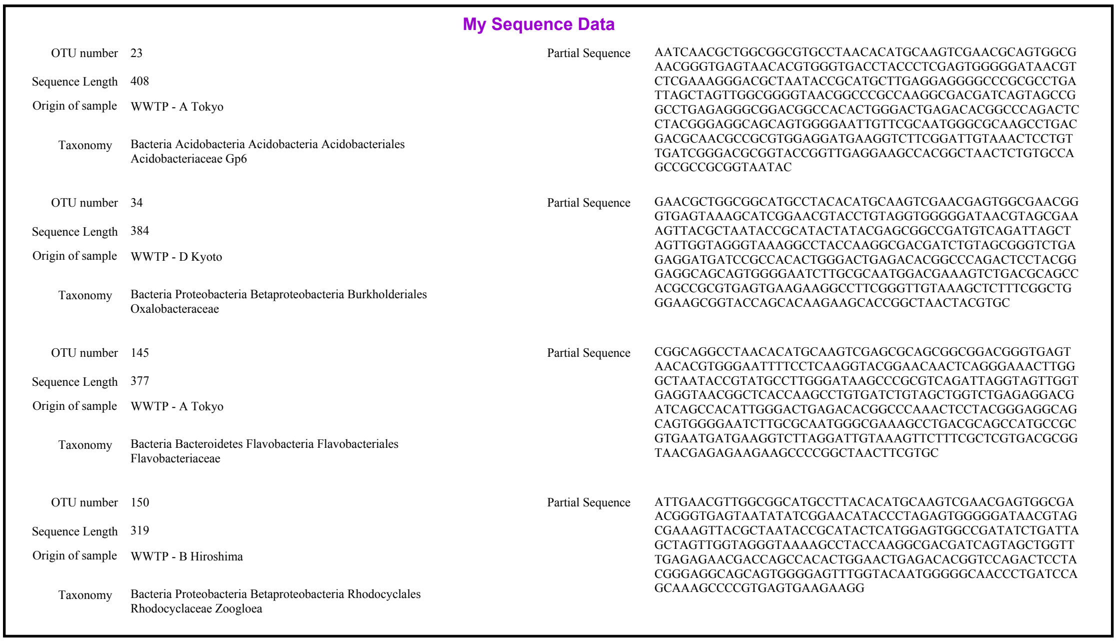


Fig.2.11 Connecting two data tables with a common field

Then, after connecting twp table user can connect the data from table 1 and 2.

For, example: If user need to see the taxonomy of OTU number 150, no need to moved to table 2. By making a simple data layout (My Sequence) as shown in Fig. 2.12, user can directly see all the details related to each OTU number.



My Sequence Data

OTU number 23
Sequence Length 408
Origin of sample WWTP - A Tokyo
Taxonomy Bacteria Acidobacteria Acidobacteria Acidobacteriales Acidobacteriaceae Gp6

Partial Sequence
AATCAACGCTGGCGGCGTGCCTAACACATGCAAGTCGAACGCAGTGGCG
AACGGGTGAGTAACACGTGGGTGACCTACCCTCGAGTGGGGGATAACGT
CTCGAAAGGGACGCTAATACCGCATGCTTGAGGAGGGGCCCGCGCCTGA
TTAGCTAGTTGGCGGGGTAACGGCCCGCCAAGGCGACGATCAGTAGCCG
GCCTGAGAGGGCGGACGGCCACACTGGGACTGAGACACGGCCCAGACTC
CTACGGGAGGCAGCAGTGGGGAATTGTTCGCAATGGGCGCAAGCCTGAC
GACGCAACGCCGCGTGGAGGATGAAGGTCTTCGGATTGTAAACTCCTGT
TGATCGGGACGCGGTACCGGTTGAGGAAGCCACGGCTAACTCTGTGCCA
GCCGCCGCGGTAATAC

OTU number 34
Sequence Length 384
Origin of sample WWTP - D Kyoto
Taxonomy Bacteria Proteobacteria Betaproteobacteria Burkholderiales Oxalobacteraceae

Partial Sequence
GAACGCTGGCGGCATGCCTACACATGCAAGTCGAACGAGTGGCGAACGG
GTGAGTAAAGCATCGGAACGTACCTGTAGGTGGGGGATAACGTAGCGAA
AGTTACGCTAATACCGCATACTATACGAGCGGCCGATGTCAGATTAGCT
AGTTGGTAGGGTAAAGGCCTACCAAGGCGACGATCTGTAGCGGGTCTGA
GAGGATGATCCGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGG
GAGGCAGCAGTGGGGAATCTTGCGCAATGGACGAAAGTCTGACGCAGCC
ACGCCGCGTGAGTGAAGAAGGCCTTCGGGTTGTAAAGCTCTTTCGGCTG
GGAAGCGGTACCAGCACAAGAAGCACCGGCTAACTACGTGC

OTU number 145
Sequence Length 377
Origin of sample WWTP - A Tokyo
Taxonomy Bacteria Bacteroidetes Flavobacteria Flavobacteriales Flavobacteriaceae

Partial Sequence
CGGCAGGCCTAACACATGCAAGTCGAGCGCAGCGGCGGACGGGTGAGT
AACACGTGGGAATTTTCCTCAAGGTACGGAACAACTCAGGGAAACTTGG
GCTAATACCGTATGCCTTGGGATAAGCCCGCGTCAGATTAGGTAGTTGGT
GAGGTAACGGCTCACCAAGCCTGTGATCTGTAGCTGGTCTGAGAGGACG
ATCAGCCACATTGGGACTGAGACACGGCCCAAACTCCTACGGGAGGCAG
CAGTGGGGAATCTTGCGCAATGGGCGAAAGCCTGACGCAGCCATGCCGC
GTGAATGATGAAGGTCTTAGGATTGTAAAGTTCTTTCGCTCGTGACGCGG
TAACGAGAGAAGAAGCCCCGGCTAACTTCGTGC

OTU number 150
Sequence Length 319
Origin of sample WWTP - B Hiroshima
Taxonomy Bacteria Proteobacteria Betaproteobacteria Rhodocyclales Rhodocyclaceae Zoogloea

Partial Sequence
ATTGAACGTTGGCGGCATGCCTTACACATGCAAGTCGAACGAGTGGCGA
ACGGGTGAGTAATATATCGGAACATACCCTAGAGTGGGGGATAACGTAG
CGAAAGTTACGCTAATACCGCATACTCATGGAGTGGCCGATATCTGATTA
GCTAGTTGGTAGGGTAAAAGCCTACCAAGGCGACGATCAGTAGCTGGTT
TGAGAGAACGACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTA
CGGGAGGCAGCAGTGGGGAGTTTGGTACAATGGGGGCAACCCTGATCCA
GCAAAGCCCCGTGAGTGAAGAAGG

Fig.2.12 Layout showing combined data from Table 1 and 2

Software used for this type of grouping called Relational Database Management Systems (RDBMS) and the relational database simply refers to the software itself.

FileMaker is such Relational Database management software and the example explained here was developed using FileMaker. Other than data combination, relational databases allow mathematical coding for calculations, arrangement or user-friendly layouts and web sharing of databases as well.

Thus, in present study FileMaker used as the relational database for management and interpretation of DNA sequence obtained from 16S rRNA pyrosequencing.