Summary Report

Training Course for Microbial Data Analysis

WFCC-MIRCEN World Data Center for Microorganism (WDCM) Beijing, China

November 23-30, 2018

By

Nattakorn Kuncharoen
(B. Sc. Ind Microbiol, Ph. D. candidate in Pharm Chem Nat Prod)

Bioresource Center, Faculty of Pharmaceutical Sciences
Chulalongkorn University (PCU)
254 Phayathai Road, Wangmai, Pathumwan District,
Bangkok 10330, Thailand
Personal Introduction

I am Nattakorn Kuncharoen. I was born on 3 February 1992 in Chachoengsao province, Thailand. I got my B. Sc. (Industrial microbiology) in 2013 and now I am studying a last year of the Ph. D. course (Pharmaceutical Chemistry and Natural Products) at the Faculty of Pharmaceutical Sciences, Chulalongkorn University, Bangkok, Thailand by the Royal Golden Jubilee Ph. D. Program under the Thailand Research Fund (TRF) 2015. My recent research is working on a taxonomy of endophytic actinomycetes and Gram-negative bacteria, microbial secondary metabolites, plants-microbes interaction, and bacterial genomes sequencing. I have experimental experiences in Kitasato University (6 months) and the Japan Collection of Microorganism (3 months).

_Culture Collection:_ Pharmaceutical Sciences Chulalongkorn University (PCU) Bioresource Center

ABSTRACT

The GCM 2.0 symposium on type strains sequencing project and the training course for microbial data analysis held by WFCC-MIRCEN World Data Center for Microorganisms (WDCM), Institute of Microbiology, Chinese Academy of Science (IMCAS) supported by Bureau of International Co-operation Chinese Academy of Science and World Federation for Culture Collections. Fourteen
WDCM-Training Course for Microbial Data Analysis

researchers from some culture collections in developing countries around the world including Thailand, United States of America, Australia, New Zealand, England, Mexico, Taiwan, Japan, Hungary, and Mongolia participated in this training course has two main purposes for discussion and promoting knowledge on microbial genomes and data analysis. In addition, the training aimed to update information of the big data center relied on the GCM 2.0 project including on-line catalogue, the SOPs for DNA extraction, sample submission, sequencing, data processing, data record, data storage and how to use the data. In conclusion, the relatively prolonged period of this training course provides participants with the opportunity of communication between scientific groups in different countries and recognition of their weakness and strength for promoting partnerships, joint projects, and the connection between microbial resource centers.

Keywords: GCM 2.0, WDCM, Culture collection, Type strain sequencing, Microbial data analysis, Training course

Topics in the training course

- Combination of Lab & Paper work: Culture Collection’s management
- Genome Sequencing and Data Analyses in Ustilaginomycotina
- Genome Sequencing and Genome Analysis Report
- Introduction of INSDC, data submission and genome annotation
- The BIG Data Center: from deposition to integration to translation
WDCM-Training Course for Microbial Data Analysis

- Genomic Data Analysis: Assembly and Annotation
- Microbiome Statistical Analysis based on Next Generation Sequencing Techniques
- Applying Advanced Bioinformatic tools and Machine Learning to Big Data Problems in Microbiomes Research
- Characterization of Bacterial Strains for Taxonomic Purposes
- Diversity of Yeast Community Isolated from Crater Lakes, Plant Leaves and Soils, and Proposal of Novel Species and Genera
- Bioinformatics: Genome Comparative Analysis and Genome Distance Calculation
- The Origin, Domestication and Genome Evolution of Saccharomyces Yeasts
- Droplet Microfluidics for High Throughput Cultivation, Screening and Sequencing of Environmental Microorganisms
- What have we learned from Clostridium difficile? – Bioinformatics in Clinical Detection
- Computational Design of Industrial Enzymes
- Microbial Defense Systems: from CRISPR to More
- WDCM Database Resources: Global Catalogue of Microorganisms and ABC Database
1. Brief Introduction of PCU

Thailand is well-known as the richest of biodiversity in south-east Asia. It contains lots of organisms including animals, plants, and microorganisms. Thailand has numerous novel species of microorganisms consisting large number of bacteria, filamentous fungi, yeasts, and mushrooms isolated from various sources such as freshwater, marine samples, insects, foods, and plants; therefore, the microbial culture collections were conducted.

The Pharmaceutical Sciences Chulalongkorn University (PCU) (Figure 1) was a bioresource center in the Faculty of Pharmaceutical Sciences, Chulalongkorn University, Bangkok, Thailand established by Prof. Dr. Somboon Tanasupawat in 1994. PCU aims to isolate, preserve, and identify new and known species of bacteria (Figure 2), provides training courses on microbial taxonomy and microbial products to students, researchers and interested people every year (Figure 3). Furthermore, PCU has a joint project called “Little Microbiologists” with the Demonstrated Elementary School of Chulalongkorn University for giving the microbiological knowledge to the students in grade 5 and 6 (Figure 4). The PCU maintains approximately 800 available strains as of Nov 2018 including novel bacterial species 90 strains, other bacteria and yeasts 200 strains, and isolated bacteria more than 500 strains. All strains held at PCU are limited to those classified in Risk Group 1 or 2.
Figure 1 Pharmaceutical Sciences Chulalongkorn University (PCU) Bioresources, 4th floor
Department of Biochemistry and Microbiology, Pharmaceutical Sciences Building

Figure 2 Isolation (a), Identification (b) and Preservation (c)
PCU is working to continuously exploit new microbial resources and to describe novel microbial taxa. We discovered lots of novel bacterial group including lactic acid bacteria, halophilic bacteria, and actinomycetes (Table 1).
Currently, PCU start working on type strains whole genome sequencing by collaborating with the Japan Collection of Microorganism (JCM), Japan and will cooperate with the Institute of Microbiology, Chinese Academy of Sciences in within 2018.

**Table 1** Novel species discovered by PCU’s researchers

<table>
<thead>
<tr>
<th>Novel species</th>
<th>Strain no.</th>
<th>Discovered person</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. shoreicorticis</em></td>
<td>MK21-7&lt;sup&gt;T&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Actinomadura montaniterrae</em></td>
<td>CYP1-1B&lt;sup&gt;T&lt;/sup&gt;</td>
<td>Songsumanus et al.</td>
<td>Int J Syst Evol Microbiol. 2016 Sep;66(9):3310-16</td>
</tr>
</tbody>
</table>
WDCM-Training Course for Microbial Data Analysis

We have many collaborators in Thailand and many countries around the world composing of the Microbes Division (JCM, Japan), Biological Resources Center (NBRC, Japan), KCTC (KRIBB, Korea), TISTR (Thailand), BIOTEC (Thailand), Kitasato University (Japan), Mahidol University (Nakhonsawan Campus, Thailand), Department of Biotechnology, Burapha University (Thailand), Department of Biology, Ramkhamhang University (Thailand), Tokyo University of Agriculture (Japan), and University of Tokyo (Japan). In conclusion, the PCU is still welcome new collaborators to join our projects in the future.

2. Benefits from the training course

The training course for microbial data analysis gave the information based on six majors purposes below:

1. World Data Center of Microorganisms (WDCM): GCM 2.0 management and its pipeline for type strains sequencing

2. Combination of Lab & Paper work: Culture Collection’s management

3. Genome Sequencing and Genome Analysis Report

4. Introduction of INSDC, data submission and genome annotation

5. The BIG Data Center: from deposition to integration to translation

6. Genomic Data Analysis: Assembly and Annotation

7. Microbiome analysis based on 16S rRNA/Metagenomic matrices
2.1 World Data Center of Microorganisms (WDCM): GCM 2.0 management and its pipeline for type strains sequencing

The WDCM was created in 1966 as the data center of the World Federation for Culture Collections (WFCC) and Microbial Resource Center UNESCO (MIRCEN). Currently, WDCM aims to provide integrated information services using big data technology for microbial resource centers and microbiologists around the world. In addition, WDCM purposes to provide an informative system to culture collections and general users of microbial resources by constructing a comprehensive data platform and several databases. This platform called “Analyzer of Bioresource Citation (ABC)” which is a data mining tool to extract information submitted from culture collection and implemented as the reference database (Figure 5).

![Figure 5 System overview of WDCM (Wu, 2018)](image)

This year, WDCM established the GCM 2.0 project for leading the sequencing of all type strains from culture collections around the world. The GCM 2.0 is aiming to finish to sequence 10,000 selected type strains of bacteria,
archaea, and fungi within 5 years (2018-2023), release the results publicly, and published the results with the depositors and bioresource centers. In my opinion, this project is very useful for many people who interesting and need to get accession.

2.2 Combination of Lab & Paper work: Culture Collection’s management

At Present, culture collections around the world are working based on the Nagoya Protocol which is an international agreement aiming at sharing of benefit arising from the utilization of genetic resources and the conventional of biological diversity in the fair and equitable way. Based on Nagoya Protocol, it focused on 2 main purposes:

(a) Quality control of strains (QC check) – Provide the standard process for the bioresource centers. The process contains viability, purity, and corrected identification (Figure 6).

(b) Minimum data set – The minimum requirements are depended on 5 questions under the Regulation EU 511/2014 including what (sample), when (time of sampling in situ), where (place of sampling, GPS), from whom (collectors/depositors), and to whom (users, customers) (Figure 7).
2.3 Genome Sequencing and Genome Analysis Report

Genome sequencing is a method to determine the complete DNA sequence of organisms at a single-time. The genome sequencing has the standard pipeline including Data sequencing, Data analysis (quality control check), Assembly, Genome annotation and Genes prediction. In the genome analysis, it has main important steps below:
WDCM-Training Course for Microbial Data Analysis

a) Quality of raw reads – raw reads data were filtered and cleaned using *Phred* quality scores for selecting only the good reads.

b) Assembly – the good raw reads were assembly and checked contamination based on K-mer depth values and GC content (mol%), afterwards the reads will be processed to contigs and scaffolds (Figure 8) which has two significant values including N50 length and GC depth for consideration.

After assembly, the genomes were annotation for predicting the genome components involving structural genes and functional genes by using some software and web-based tools, for example, PGAP, DFAST, MiGAP, VFDB, KEGG, COG, T3DB, and ARDB.

**Figure 8** Assembly process
WDCM-Training Course for Microbial Data Analysis

2.4 Introduction of INSDC, data submission and genome annotation

In this topic, we learned about the overview of INSDC (Figure 9), how to submit the genome sequences data into the DDBJ, how to submit raw data in the fastq format to the Sequence Read Archive (SRA), how to annotate the genomes using a web-based tool named “DFAST” which created by Tanizawa et al. (2016), and the detail in the GenBank format (Figure 10).

Figure 9 Overview of INSDC

Figure 10 GenBank data format of genome sequences
2.5 The BIG Data Center: from deposition to integration to translation

BIG, Beijing Institute of Genomics, which is a part of the Chinese Academy of Sciences. This institute created and collected lots of genomic data of humans, animals, plants, and microbes. The institute has published many publications per year. Recently, BIG will become the big data center of the world like the NCBI, DDBJ or EMBL in the future. Moreover, BIG launched a plenty of genome projects, for instance, ScienceWikis, EWA Atlas, iDOG-10K genomes of domestic dogs, and Global Biodiversity and Health Big Data (BHBD) (Figure 11).

2.6 Genomic Data Analysis: Assembly and Annotation

The issue described about the data analysis including assembly and annotation process. In the data analysis, there has many bioinformatic tools using for assembly such as QIME2 and SPAdes, and for annotation involving PGAP (NCBI pipeline), DFAST (DDBJ pipeline), and MiGAP (DOE-JGI pipeline). However, Dr. Wenyu Shi and WDCM’s team have created a web-based tool for
WDCM-Training Course for Microbial Data Analysis

assembly, annotation, and metagenomic analysis named “gcMeta”
(https://gcmeta.wdcm.org/) which users are easy to use and operate following the
manual (Figure 12). In addition, this topic provided a web-based tools using for
gene prediction consisting of KEGG, MetaCyc, Pfam, Rfam, and PHI.

![Figure 12 gcMeta – a web-based tool](image)

2.7 Microbiome analysis based on 16S rRNA/Metagenomic matrices

This session provided a concept of microbiome analysis based on
mathematical models and statistics. It divided the analysis into 2 types including
Alfa-diversity that observed a distribution of taxa in the sample, and Beta-
diversity which aimed to study the relationship of the microbial community. The
study used “R-program” for analysis.
WDCM-Training Course for Microbial Data Analysis

3. Suggestion on WDCM work

From my point of view, Firstly WDCM should improve and update the website to present every time because users might get incorrect data. Secondly, WDCM should provide budget to the researchers from the collection cultures around the world for working short-time about genome sequencing in IMCAS, BGI, or BIG. Finally, WDCM should support research funding for collaborative projects between China and other countries.

4. Comments or suggestions on the training course

In my opinion, the training course’s schedule should be exactly rearranged in accordance with the first time that WDCM sending to all participants. The course should focus on “Microbial data analysis-based on genome sequences” as official schedule. It might be perfect if the training course could have more related training’s issues and practices on microbial genomic data analysis, for instance, the participants must begin with the genomic DNA extraction, sequentially by genome sequencing with practical platforms, genome data analysis, genome assembly, genome annotation, genome submission, and how to use the genome for taxonomic purposes. Furthermore, the training host should separate the course for bacterial and fungal taxonomists because they are different in that fields. Finally, may I said that a lot of topics in the course look like the IMCAS’s presentation.
5. Suggestion on further cooperation between WDCM and your collection

Possible collaborative projects:

“Sequencing of Novel Species Isolated from Thai Bioresources, Plants-Microbes interaction, and Microbes-associated with Foods using microbiome data analysis”

Rationale and Reasons:

Thailand is a relatively small country in the south-east asia, but she is rich in biodiversity, containing approximately 8-10 % of the total species known thus far. Thailand has numerous sources that interesting to study including animals, plants, foods, marine and mangrove ecosystems, and tropical forests. It is believed that more than 100,000 species of living microorganisms await discovery. Therefore, Thailand is situated in one of the richest areas of the world regarding biological resources and microbial biodiversity.

Our study was investigated to identify novel and useful bacteria isolated from variety of Thai bioresources based on polyphasic taxonomic characteristics and whole genome sequences analysis. Also, we conducted this research to understand bacterial activities: production of bioactive compounds, production of industrial enzymes, plants-growth promoting agents, start culture for fermented foods, and production of bio-drugs using cloning technique. In addition, the study aimed to understand the microbial community of microbes associated with plants and foods in Thailand using the microbiome data analysis.
The findings of the present study may shed light on the biodiversity of new and useful bacterial species recovered from bioresources in Thailand. Moreover, the study may provide the genomic data which can be used for the discovery and development of microbial products in the future.