TSPE15020-005-12

# 균류 LSU 표준염기서열 DB기반 미생물다양성 및 환경유전체 분석기술 개발

Development of microbial diversity and environmental genome analysis method based on fungal LSU sequence DB



한국생명공학연구원

# 제 출 문

극지연구소장 귀하

본 보고서를 “기후변화에 의한 킹조지섬 생태계 변화 연구” 과제의 위탁연구 “균류 LSU 표준염기서열 DB기반 미생물다양성 및 환경유전체 분석기술 개발에 관한 연구” 과제의 최종보고서 (“2015년_최종보고서_KRIBB”)로 제출합니다.



2016. 1. 18

(본과제) 총괄연구책임자 : 홍 순 규

위탁연구기관명 : 한국생명공학연구원

위탁연구책임자 : 김 경 모

위탁참여연구원 : 오 정 수

“ : 최 한 나

“ : 김 선 규

# 요 약 문

Ⅰ. 제 목

균류 LSU 표준염기서열 DB기반 미생물다양성 및 환경유전체 분석기술 개발

Ⅱ. 연구개발의 목적 및 필요성

<연구 필요성>

- 미생물 다양성 및 생태의 생물정보학적 연구는 기후변화, 농업, 의료, 환경 등 인간생활에 밀접한 영향을 미치며 다양한 산업분야 전반에 높은 활용도를 가짐
- 이미 기존에 454 데이터 기반 16s rRNA 기반 미생물 다양성 분석 파이프라인은 구축되어 있으나 주어진 환경에서의 미생물의 역할과 환경내의 기능은 분석하지 못함.
- 따라서 환경 내 존재하는 모든 유전자 서열을 시퀀싱하는 whole genome shotgun 메타지놈 데이터를 분석할 수 있는 파이프라인이 필요함.
- NGS기술의 지속적인 발전으로 인해 대용량 염기서열의 분석필요성이 지속적으로 증가함에 따라 분산처리가 가능한 염기서열 클러스터링 프로그램이 필요함.
- 이에 본 연구팀은 In-memory 기반 대용량 염기서열 클러스터링 프로그램을 개발하였음. 그러나 개발된 시스템은 안정성과 실행 편의성의 부족으로 일반 사용자들이 활용하기에 어려운 점이 있기 때문에 고도화가 필요함.

<연구목적>

- Whole genome shotgun 메타전사체 데이터를 분석 할 수 있는 프로그램을 개발하여 생태계 융합연구를 위한 생물정보학적 분석 인프라 제공
- 대용량 염기서열을 클러스터링 할 수 있는 기 개발된 in-memory 기반 염기서열 클러스터링 프로그램의 고도화를 진행하여 고가용성과 사용성을 개선하여 활용성을 높이는데 기여.

Ⅲ. 연구개발의 내용 및 범위

- Whole genome shotgun 메타전사체 데이터를 분석 파이프라인 구축
- NGS 기반 미생물다양성 분석 프로그램 개발 및 균류 LSU 염기서열 DB와의 연동
- In-memory 기반 염기서열 클러스터링 프로그램 고도화

Ⅳ. 연구개발결과

- 정확성 높은 메타지놈 어셈블리 프로그램을 활용한 자동화된 메타지놈 염기서열 assembly 파이프라인 구축을 완료.
- 메타지놈 유전자 예측 및 기능 annotation 을 위한 분석 파이프라인 구축 완료

- 대용량 염기서열 클러스터링을 클라우드 환경에서 손쉽게 구동할 수 있도록 설정 파일 간소화 및 쉘 스크립트 구현 완료
- 기존에 개발된 프로그램 보다 실행속도가 더 빨라지도록 추가적인 기능 구현 완료
- 기존 CLUSTOM의 문제점과 처리속도 향상을 높인 싱글 노드에서 구동 가능한 Java version의 CLUSTOM 개발 완료
- 기존에 개발된 프로그램의 고가용성과 데이터 안정성 개선을 위한 아키텍쳐 개선 완료

Ⅴ. 연구개발결과의 활용계획
- 남극 토양 균류 다양성 및 생태연구에 기여
- 미생물 다양성 분석 프로그램과의 연동을 통해 Whole genome shotgun 메타지놈에 대한 상용서비스 제공 가능
- 대용량 염기서열의 데이터에 대한 안정적이고 편리한 분석 가능

# Summary

Ⅰ. Title

Development of microbial diversity and environmental genome analysis method based on fungal LSU sequence DB

Ⅱ. Research goal and background

<Research background>
- Bioinformatics studies on microbial diversity and ecology can contribute to diverse related fileds including climate change, agriculture, medical industry and environments.
- While we already developed the bioinformatics pipeline for analyzing 16S rRNA pyrosequencing data of microbial diversity, survey of microbial roles under a given environment is not conducted at the gene level.
- Consequently, it is necessary to develop a pipeline that enables us to analyze whole genome shotgun metagenome sequence data.
- On behalf of dramatic advance of next generation sequencing technologies, it is necessary to develop a sequence could clustering program that can analyze large-scale metagenomic sequence data.
- We thus developed a cloud-based in-memory sequence clustering program. However, software update for user's friendly interface should be further conducted in terms of system stability and user's convenience.

<Research goal>
- Software development for analyzing NGS sequences from microbial community and whole genome shotgun metagenome studies.
- By developing bioinformatics programs that analyze large-scale sequence data, provide bioinformatics infrastructure that facilitates studies of microbial diversity and ecology.

Ⅲ. Research content and scope
- Developing a bioinforamtics pipeline for analyzing whole genome shotgun metagenome sequence data

- Software development for NGS sequence datasets and its incorporation into the fungal LSU sequence database
- CLOUD-based software development for conducting ribosomal RNA sequence clustering

Ⅳ. Research results
- It was completed that the bioinformatics pipeline can analyze whole genome shotgun metagenome sequence data.
- Developed a pipeline, which can annotate structural and functional features for microbial genes obtained from metagenome sequence data.
- Finished bioinformatics automation, which enables us to easily analyze metagenomic DNA sequences under a cloud-computing environment.
- Updating the clustering software in term of computing speed by optimizing the software source codes.
- By using the In-Memory Data Grid method, we upgraded the CLUSTOM program that enables to cluster the large sequence data without lack of system memory resource.
- In order to calculate the large sequence data clustering, the parallel system was developed.



Ⅴ. Plan for research application
- Completing bioinforamtics infrastructure to contribute sequence-based analysis on data obtained from Antarctic microbial samples
- Providing a bioinformatics service with an aid of softwares we developed
- Making possible to user-friendly analyze large-scale metagenome sequence data

# Contents

# 목 차

# 제 1 장 서론

## 제 1-1절: In-memory 기반 염기서열 클러스터링 클라우드 환경 구축 고도화 필요성

1. 차세대 염기서열 결정법의 지속적인 발전으로 더 싼값으로 빠르게 많은 양의 서열을 생산하는 것이 가능해 짐에 따라 엄청난 양의 16s rRNA서열 데이터 대한 분석 필요성이 대두됨.
2. 또한, 이러한 대규모 염기서열 데이터를 분석하는데 필요한 대규모 전산자원의 부족 및 기존의 전산학적 접근 방법의 한계가 보이기 시작하기 때문에 클라우드 환경을 지원하며 빠르게 처리 가능한 새로운 전산기술의 적용이 필요함.
3. 이에 따라 본 연구팀은 작년에 극지연구소 위탁과제를 수행하면서 In-Memory Data Grid 기반 대용량 염기서열 클러스터링의 개발을 성공적으로 완료하였음.
4. 그러나 개발된 시스템은 안정성이 떨어지거나 컴퓨터를 잘 다루지 못하는 일반 사용자들이 일반 분산환경 및 클라우드 환경에서의 실행에 어려운 점이 있었음.
5. 따라서, 기존의 개발된 시스템의 성능 및 안정성을 높이고 편리하고 쉬운 실행이 가능하도록 고도화하여 활용성을 높이는 것이 필요함.



## 제 1-2절: Whole 메타전사체 shotgun 염기서열 분석 파이프라인 구축의 필요성

1. 단일 유전자 기반(예.16s rRNA)의 메타지놈 샘플에 대한 미생물 다양성 분석은 본 연구팀이 자체 개발한 파이프라인 뿐만 아니라 MOTHUR(Schloss et al. 2009), QIIME(Kuczynski et al. 2012) 와 같은 분석 파이프라인을 통해 널리 수행되고 있음.
2. 그러나 단일 유전자 접근법은 주어진 환경이 미생물 다양성 및 군집구조 분석의 수행이 가능하지만, 해당 환경에서 존재하는 미생물이 어떤 역할을 하는지, 그 미생물에 속한 유전자의 환경 내 에서의 기능은 무엇인지는 알기 어렵기 때문에 미생물과 해당 환경과의 상호작용을 이해하는데 한계가 있음.
3. 이에 따라 주어진 환경 내 존재하는 모든 유전자 서열을 시퀀싱하는 whole genome shotgun 기법을 통해 생산된 메타지놈 데이터를 분석할 수 있는 파이프라인이 필요함.
4. 현재까지 whole genome shotgun 기반의 메타지놈 데이터를 분석하기 위해 EBI Metagenomics Portal(http://www.ebi.ac.uk/metagenomics) , MG-RAST, IMG/M 등과 같은 여러 가지 파이프라인이 개발되었으나 웹상에서 실행이 가능하기 때문에 사용자들이 자신의 데이터를 원격의 서비스 웹서버로 업로드 해야 하기 때문에 대량의 메타지놈 데이터를 업로드하는데 어려움이 있을 뿐만아니라 연구자 필요에 따른 분석모듈 및 데이터베이스를 탑재하기 힘든 단점이 있음.
5. 비록 MEGAN4 (Huson et al. 2011)과 같은 GUI 환경을 갖춘 PC에 구동 가능한 툴이 있지만 대용량 데이터를 처리하기 힘들며, 위와 같은 파이프라인에 비해 다양한 분석을 제공

하고 있지 못함.

6. 따라서 로컬 서버상에서 대용량 whole genome shotgun 메타지놈 데이터를 쉽게 처리 할 수 있을 뿐만 아니라 연구자의 선호에 수정이 간편한 파이프라인이 필요함.

# 제 2 장 국내외 기술개발 현황

## 제 2-1절: 대용량 염기서열 클러스터링 프로그램 개발 현황

1. 현재까지 차세대 염기서열 결정법에 의해 생산된 대용량 염기서열 데이터를 클러스터링 할 수 있는 프로그램은 CD-HIT(Li and Godzik 2006),과 HPC-CLUST(Rodrigues et al. 2013) 등이 있음.
2. CD-HIT은 greedy heuristic clustering 기법을 사용하여 처리 속도가 매우 빠르나, 단일 노드에서의 병렬처리만을 지원하기 때문에 프로그램의 성능 및 컴퓨터의 자원 임계치를 넘어서는 대용량의 데이터에 대해서는 처리하지 못함. 무엇보다 hierarchical clustering 알고리즘에 비해 정확도가 매우 떨어지는 단점이 존재.
3. HPC-CLUST 경우 hierarchical clustering 알고리즘을 사용하여 정확성이 높고 MPI를 활용한 병렬 및 분산처리를 지원하기 때문에 속도도 빠른 장점이 있으나, 클러스터링 시 필요한 Similarity matrix데이터를 프로그램 내에서 연산을 통해 생성하는 것이 아닌 외부에서 입력받아 실행되기 때문에 단독으로 수행될 수 없음. 무엇보다 서열 유사도를 기반으로 Similarity matrix데이터를 생성하는데 엄청난 시간이 들기 때문에 대용량 데이터에 대한 클러스터링을 수행하는 것이 실질적으로 매우 어려움.
4. 본 연구팀은 이러한 기존의 서열 클러스터링 프로그램의 한계를 극복하기 위해 대용량 염기서열 클러스터링을 분산환경 및 클라우드 환경에서 실행 가능한, 메모리 기술 기반 클러스터링 분산처리 시스템을 개발하였음.
5. 본 프로그램은 기존에 본 연구팀에서 고안한 정확한 서열 클러스터링을 위한 CLUSTOM(Hwang et al, 2013) 알고리즘을 기반으로 정확한 서열 클러스터링이 가능.
6. 그러나 실행 안정성이 떨어지고 분산처리를 위한 여러 가지 설정이 필요하기 때문에 사용하기 어려운 단점이 있음

## 제 2-2절: Whole 메타전사체 shotgun 염기서열 분석 파이프라인 개발 현황

1. 최근 NGS 기술의 발전으로, 다양한 환경 샘플에 대해 싼 가격으로 deep sequencing 이 가능하게 됨에 따라, whole 메타전사체 shotgun 데이터를 분석할 수 있는 여건이 마련됨. whole 메타전사체 shotgun 데이터는 단일 유전자 기반 메타지놈 데이터에서 분석하지 못하는 해당 환경에서 존재하는 미생물의 기능과 역할을 파악할 수 있기 때문에 그 필요성이 점점 증대되고 있음.
2. 현재까지 whole genome shotgun 기반의 메타지놈 데이터를 분석하기 위한 분석 파이프라인은 유럽에서 개발된 EBI Metagenomics Portal, 미국 연구그룹에서 개발된 MG-RAST, IMG/M이 각각 웹상에서 분석파이프라인을 제공하며, MEGAN4가 GUI를 탑재한 PC 설치 프로그램을 제공하고 있음.
3. 위에서 언급된 각각의 파이프라인은 각각 파이프라인별로 입력 데이터에 대한 functional annotation에 대한 제공 정보가 다름. EBI Metagenomics Portal 의 경우 GO slims 기반

의 데이터베이스(http://www.geneontology.org/GO_slims/goslim_metagenomics.obo)에 대해서만 정보를 제공. MG-RAST는 COGs (Tatusov et al. 2003)나 KEGG (Kanehisa et al. 2000), GO (Gene Ontology Consortium. 2004) 등 다양한 데이터베이스를 기반으로 annotation 정보를 제공해주나 protein family 분석에 있어 주로 사용되는 pfam이 아닌 FIGfams (Meyer et a.l 2009)을 사용. 이러한 웹 기반의 파이프라인은 사용자가 데이터를 업로드해야 해야 하기 때문에 대용량 데이터를 업로드하기 어려우며, 사용자가 자신의 모듈이나 외부 리소스를 탑재하지 못하는 단점이 존재.

4. MEGAN4의 경우 PC 상에서 설치가 가능하나 컴퓨터 리소스 한계로 인해 대용량 데이터를 다루기 힘들며, 지원하는 데이터베이스 리소스가 적을 뿐 아니라 외부 리소스를 탑재할 수 없는 단점이 존재
5. 본 연구팀은 극지연구소와 공동으로 454 파이로시퀀싱 기반 16rRNA 미생물 다양성 분석 파이프라인을 개발하였으나 일루미나 시퀀서 기반의 whole 메타전사체 shotgun 데이터에 대한 분석 파이프라인은 개발이 미미한 실정임.

# 제 3 장 연구개발수행 내용 및 결과

## 제 3-1절: In-memory 기반 염기서열 클러스터링 클라우드 환경 구축 고도화

1. 기존 연구로 개발된 In-memory 기반 염기서열 분산처리 클러스터링 도구인 CLUSTOM-CLOUD를 분산 및 클라우드 환경에서 쉽게 구동할 수 있도록 고도화.

가. Cluster computer가 갖춰진 분산 환경이나 아마존 EC2와 같은 클라우드 환경에서 편리하고 빠르게 구동할 수 있도록 기존의 설정파일을 간소화 함

나. 구조화된 xml설정 파일에서 클라우드 환경에서 구동할 때 필요한 옵션 값만 입력하면 됨. 만약 클라우드 환경이 아닌 일반 클러스터 환경에서는 설정할 필요 없이 간편하게 사용가능하도록 함. (그림 1)

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE properties SYSTEM "http://java.sun.com/dtd/properties.dtd">
<properties>
        <comment>CLUSTOM PROPERTIES</comment>
        <entry key="imdg_server">127.0.0.1:5701</entry>
        <entry key="imdg_ip_range">127.0.0.*</entry>
        <entry key="imdg_ip_member">127.0.0.1</entry>
        <entry key="imdg_port_number">5701</entry>
        <entry key="the_number_of_threads">8</entry>
        <entry key="chunk-size">2000</entry>
        <entry key="aws-enabled">false</entry>
        <entry key="access-key"></entry>
        <entry key="secret-key"></entry>
        <entry key="region"></entry>
        <entry key="host-header"></entry>
</properties>
```

그림 1. 분산 및 클라우드 환경에서의 구동을 위한 설정 파일

다. 일반적으로 분산처리 프로그램을 구동 시 여러 설정으로 인해 사용에 어려움이 있음에 따라 손쉽게 구동할 수 있도록 구동에 필요한 설정 및 명령어를 한번에 실행하는 쉘

스크립트 파일(그림 2)을 만들어 사용자들이 손쉽게 사용할 수 있도록 함.

```
#!/usr/bin/env bash
function getPid(){
  Pid=$(ps -ef|grep kr.re.kribb.clustom.server.Server|grep -v grep|awk {'print $2'}|xargs)
  return $Pid
}
function getClientPid(){
  Pid=$(ps -ef|grep kr.re.kribb.clustom.client.Client|grep -v grep|awk {'print $2'}|xargs)
  return $Pid
}
function printServerUsage(){
  echo "Server Usage: server {start|stop}"
}
function printClientUsage(){
  echo "Client Usage: client {start|stop} {-g 0.03 -r 3000 -i fileName -d true}"
}
function getCygwinPid(){
  Pid=$(ps -ef|grep java|grep -v grep|awk {'print $2'}|xargs)
  return $Pid
}

this="${BASH_SOURCE-$0}"
common_bin=$(cd -P -- "$(dirname -- "$this")" && pwd -P)
script="$(basename -- "$this")"
this="$common_bin/$script"
# convert relative path to absolute path
config_bin=`dirname "$this"`
script=`basename "$this"`
config_bin=`cd "$config_bin"; pwd`
this="$config_bin/$script"

export CLUSTOM_PREFIX=`dirname "$this"`/..
DEFAULT_CONF_DIR="conf"
CLUSTOM_CONF_DIR="${CLUSTOM_CONF_DIR:-$CLUSTOM_PREFIX/$DEFAULT_CONF_DIR}"

if [ -f "${CLUSTOM_CONF_DIR}/clustom-env.sh" ]; then
  . "${CLUSTOM_CONF_DIR}/clustom-env.sh"
fi
#echo "[DEBUG] ClustomCloud Home Directory: ${CLUSTOM_HOME} "
```

그림 2. 구동을 위한 쉘 파일

라. 클라우드 환경에서 본 연구결과물의 안정성과 확장성을 확인하기 위해 아마존 EC2 클라우드 환경에서 실험을 진행함. 실험 환경은 A*pplication* 노드는 아마존 EC2의 High-CPU Extra Large Instance 로서 2.8 GHz 32 코어 CPU와 60GB 메모리를 탑재하였으며, *Cluster* 노드는 EC2 High-Memory Extra Large Instance로서 2.5 GHz 4 코어 CPU와 30.5 GB 메모리를 탑재함. 실험 데이터는 Human Microbiome Project(Peterson et al. 2009) 에서 랜덤으로 100만개의 서열을 추출하였음. 실험 결과 클라우드 환경에서 대용량 서열을 빠르고 안정적으로 클러스터링 가능함을 확인(표 1).

표 1. 아마존 EC2 클라우드 환경에서의 100백개 서열 클러스터링 실험 결과

| **EC2 nodes** | **1 app, 20 clusters** | **1 app, 30 clusters** | **1 app, 40 clusters** |
|---|---|---|---|
| Processor cores | 80 | 120 | 160 |
| Wall clock time | 20 h:38 m | 14 h:05 m | 11 h:34 m |
| Cluster setup time | 5 m | 5 m | 5 m |
| Reads uploading time | 36 m | 36 m | 36 m |

| | | | |
|---|---|---|---|
| *k*-mer distance calculation time | 16 h:34 m | 11 h:03 m | 8 h:50 m |
| Initial cluster determination time | 55 m | 27 m | 25 m |
| NW distance calculation time | 2 h:11 m | 1 h:38 m | 1 h:22 m |
| Final cluster determination time | 17 m | 16 m | 16 m |

2. 중복 서열 제거 기능 구현을 통한 전체 클러스터링 실행 시간 단축

가. 일반적으로 차세대 시퀀싱(NGS)을 통해 나온 대용량 메타지놈 서열에는 추출한 샘플의 미생물 다양성의 특성에 따라 많은 중복서열이 발생함. 이러한 중복서열은 클러스터링 과정에서 서열간의 거리 계산 및 유사도 계산의 시간에 많은 영향을 미침. 따라서 중복서열의 제거 기능을 구현하여 전체 클러스터링 실행시간을 단축하도록 함. 데이터의 중복 개수에 따라 최대 2배 정도 시간 단축 효과를 보임 (그림 3).

나. 입력 데이터의 미생물 다양성 complexity 에 따라 서열의 중복개수에 차이의 정도가 다르며, complexity가 낮을수록(다양성이 낮을수록) 중복개수가 많아짐. 따라서 미생물 다양성이 낮은 샘플일수록 시간 단축의 효과가 더 발휘됨.

다. CLUSTOM 알고리즘에서는 이러한 중복서열도 네트웍 상의 클러스터의 중심을 선택하는데 필요한 중요한 요소임. 따라서 서열간의 거리 및 유사도 계산이 끝난 후 클러스터링을 결정할 시 중복서열의 개수도 참조 할 수 있도록 함. 중복된 서열들은 최종 OTU 결과에서 복구되며 최종 군집화 결과는 기존의 오리지널 CLUSTOM 버전과 차이가 없도록 구현.

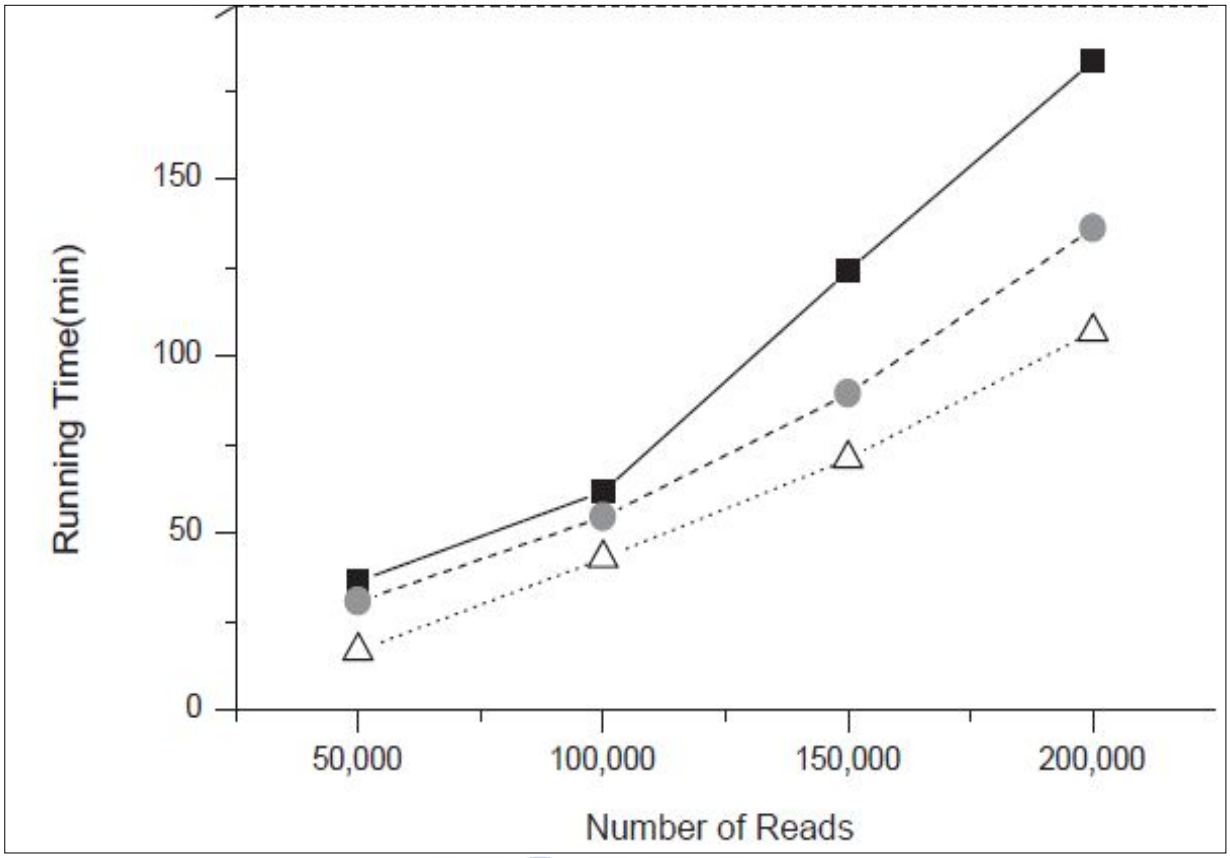


그림 3. 입력 데이터의 중복개수에 따른 실행시간 비교 그래프



3. *k*-mer 문자형 데이터를 숫자형 데이터 타입으로 변환하는 기법을 통해 서열간 거리 계산 속도 및 메모리 효율 개선

가. CLUSTOM 알고리즘에서 가장 계산시간이 많이 걸리는 부분은 각 서열간의 *k*-mer 거리 계산으로, 본 연구에서는 이를 개선하기 위해 *k*-mer 문자형을 숫자형 테이터 타입으로 바꾸는 기법을 고안하여 적용하였음 (그림 4).

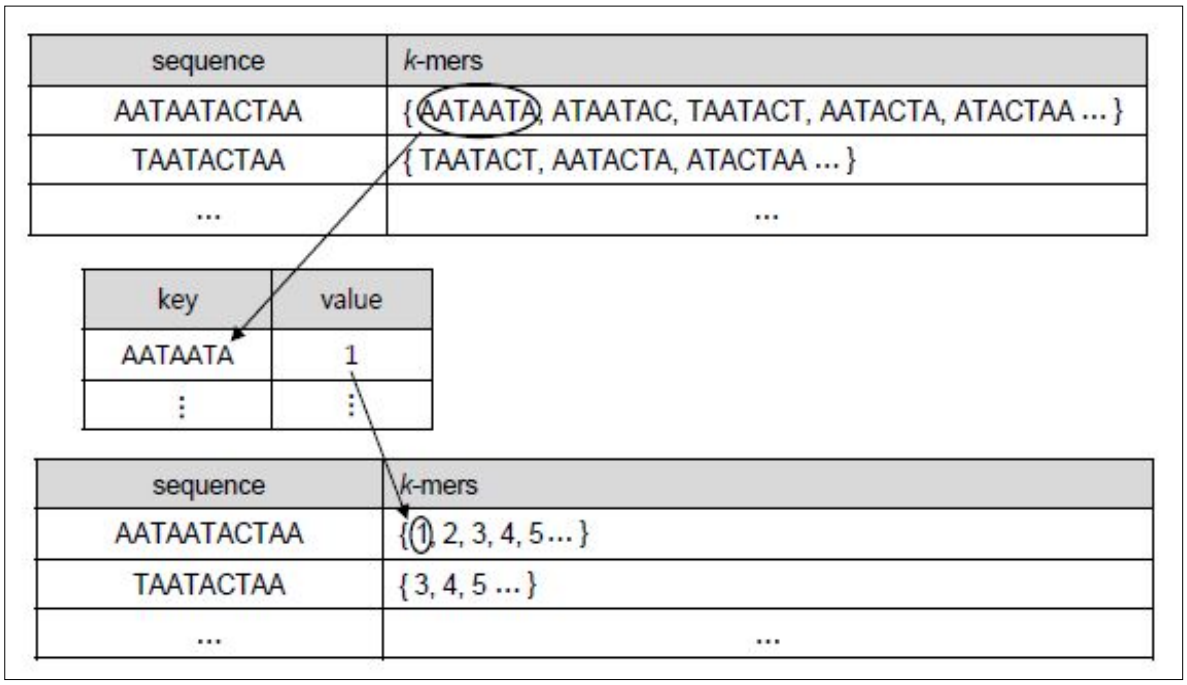

그림 4. *k*-mer 문자형을 숫자형 테이터 타입으로 변환하는 과정을 나타낸 모식도

나. 변환된 데이터 및 *k*-mer 거리 계산을 위한 정보는 In-Memory Data Grid(IMDG) 기반의 자료구조에 저장되나, 변환 과정에서 사용되는 데이터는 로컬 메모리를 사용하도록 하여 빠른 속도로 변환이 가능하도록 함.

다. 숫자형 데이터 타입에 비해 문자형 데이터 타입은 메모리를 더 점유할 뿐만 아니라 연산속도도 더 느림. 따라서 고안된 기법을 통해 적용하지 않았을 때보다 연산속도와 메모리 사용량에서 최대 2배이상 높은 효율을 보임(그림 5).

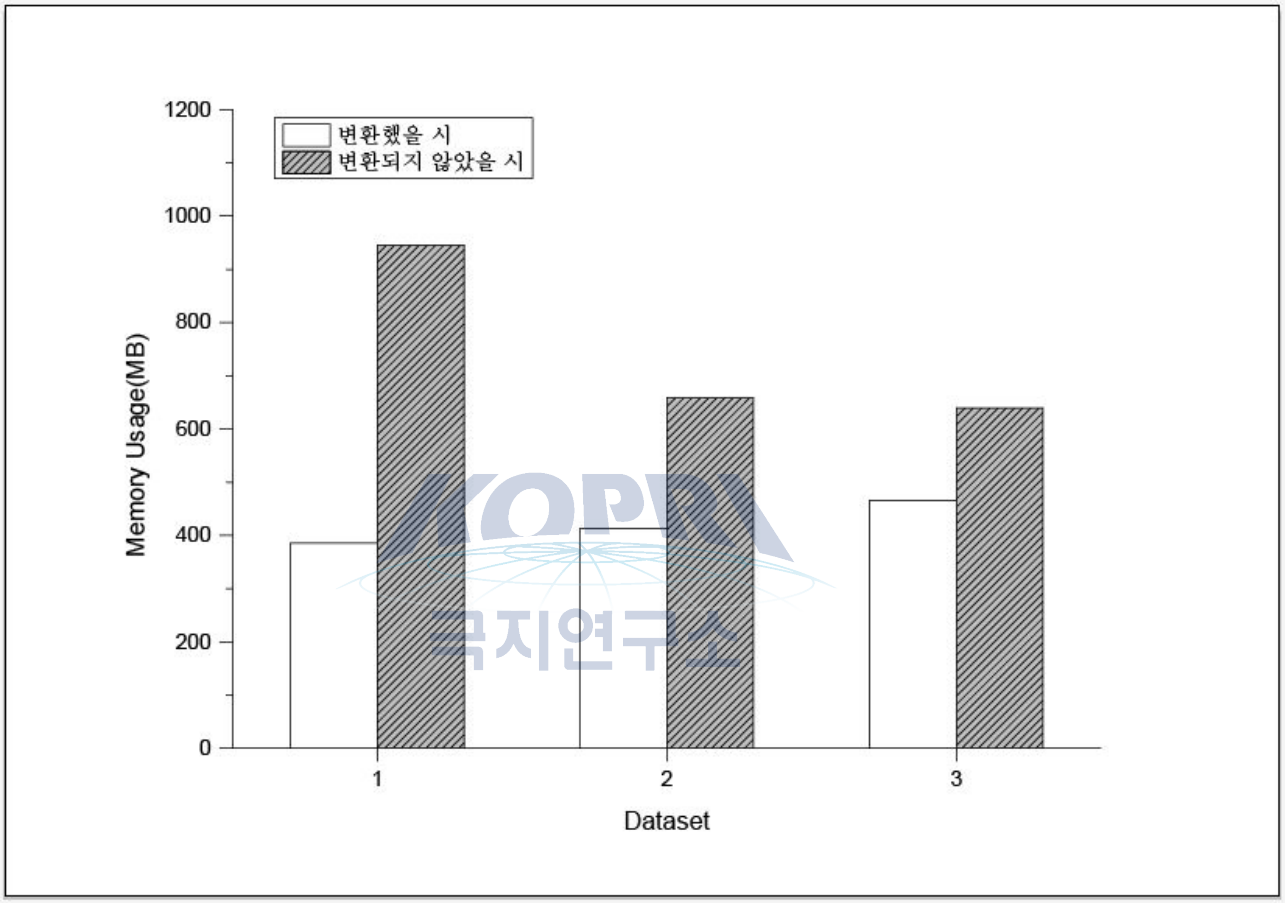


그림 5. *k*-mer 문자형 데이터를 숫자형 데이터 타입으로 변환했을 때와 안했을 때의 메모리 사용량 그래프.

라. 특히, 이 기법을 적용하여 데이터 타입을 변환하지 않았을 때는 입력 데이터의 complexity 특성에 따라 메모리의 사용량이 많은 차이가 있었으나, 이 기법을 적용하여 데이터 타입을 변환 했을 시는 일정하면서도 적은 메모리 사용량을 보임. 이에 따라 안정적인 *k*-mer 거리 계산이 가능하도록 함.

4. 싱글 노드에서 구동 가능한 Java version의 CLUSTOM개발

가. 기존의 개발된 CLUSTOM은 다른 염기서열 클러스터링 프로그램에 비해 정확도가 우수하지만 대용량 서열의 경우 많은 메모리가 필요되기 때문에 종종 메모리 부족문제가 발생함. 또한 C언어로 구현되었기 때문에 윈도우 운영체제를 지원하지 않으며 gcc

4.12 이하 버전에서만 실행이 가능하기 때문에 사용이 매우 제약적임. 이에 본 연구팀에서는 싱글노드에서 구동 가능한 Java 버전의 CLUSTOM을 개발.

나. 개발된 버전을 운영체제에 독립적이며 특별한 세팅과정 없이 바로 실행이 가능.

다. 또한 메모리 자원에 최적화 되어 설계 및 구현 되었기 때문에 기존 버전의 메모리 부족문제를 해결하여 대용량 서열에 대해서도 적은 메모리를 가지고 구동이 가능. 기존 버전은 데이터의 complexity 특성에 따라 메모리 사용량의 차이가 크고 complexity 매우 낮은 경우에 대해 급격한 메로 사용량의 증가로 인해 메모리 부족 에러가 나타나는 문제가 존재. 새로 개발된 Java 버전은 기존 버전과 달리 데이터의 complexity 특성에 따라 메모리 사용량이 거의 일정하여 안정적으로 구동 가능 (그림 6).

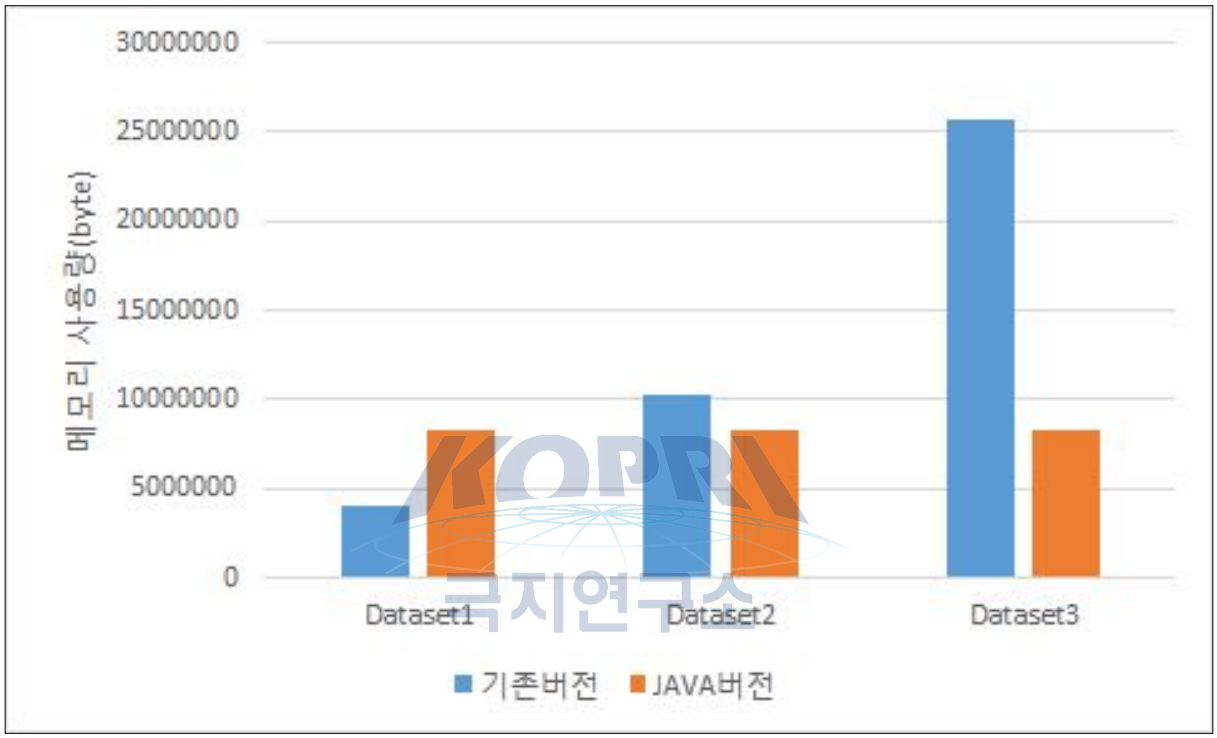


그림 6. 기존 버전과 Java 버전과의 메모리 사용량 비교

라. 최적화된 구현 및 멀티 스레드 도입으로 기존 버전에 비해 비약적인 속도 향상을 이룸(그림 7). 특히 데이터의 complexity 특성에 따라 최대 5배 이상 빠른 처리 속도의 향상을 보임.

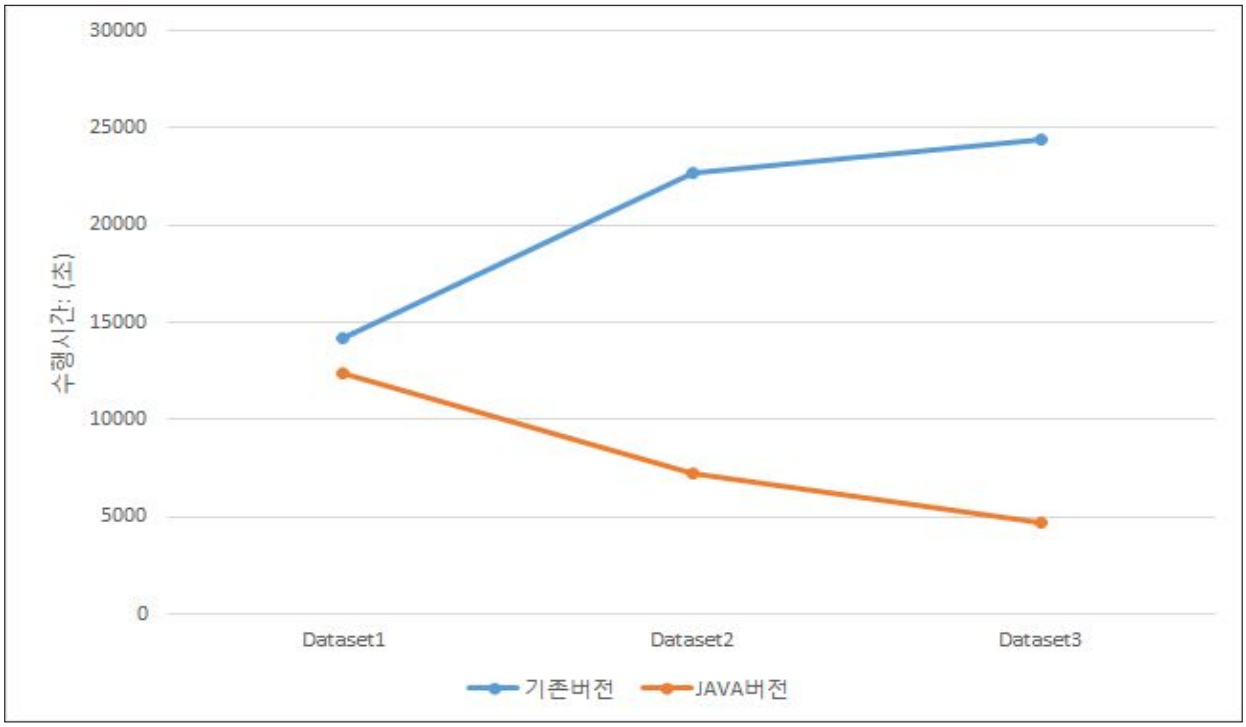


그림 7. 기존 버전과 Java 버전과의 처리 속도 비교

5. IMDG 기반 염기서열 클러스터링 프로그램의 고가용성과 데이터 안정성 개선을 위한 아키텍쳐 개선

가. 기존 IMDG 기반 염기서열 클러스터링 프로그램의 분산 및 클라우드 환경에서의 구동 시 고가용성과 데이터의 안정성을 높이기 위해 새로운 아키텍처 도입.

나. 분산 환경 상에서 높은 고가용성과 확장성을 보장하고 대규모 서열 데이터를 IMDG 상에서 저장 처리하기 위해, 본 연구팀은 기존의 클라이언트 서버 아키텍쳐에서 하이브리드 아키텍처로 변경(그림 8). 개선된 아키텍처는 클러스터(*Cluster*)와 어플리케이션 (*Application*) 으로 나뉨.

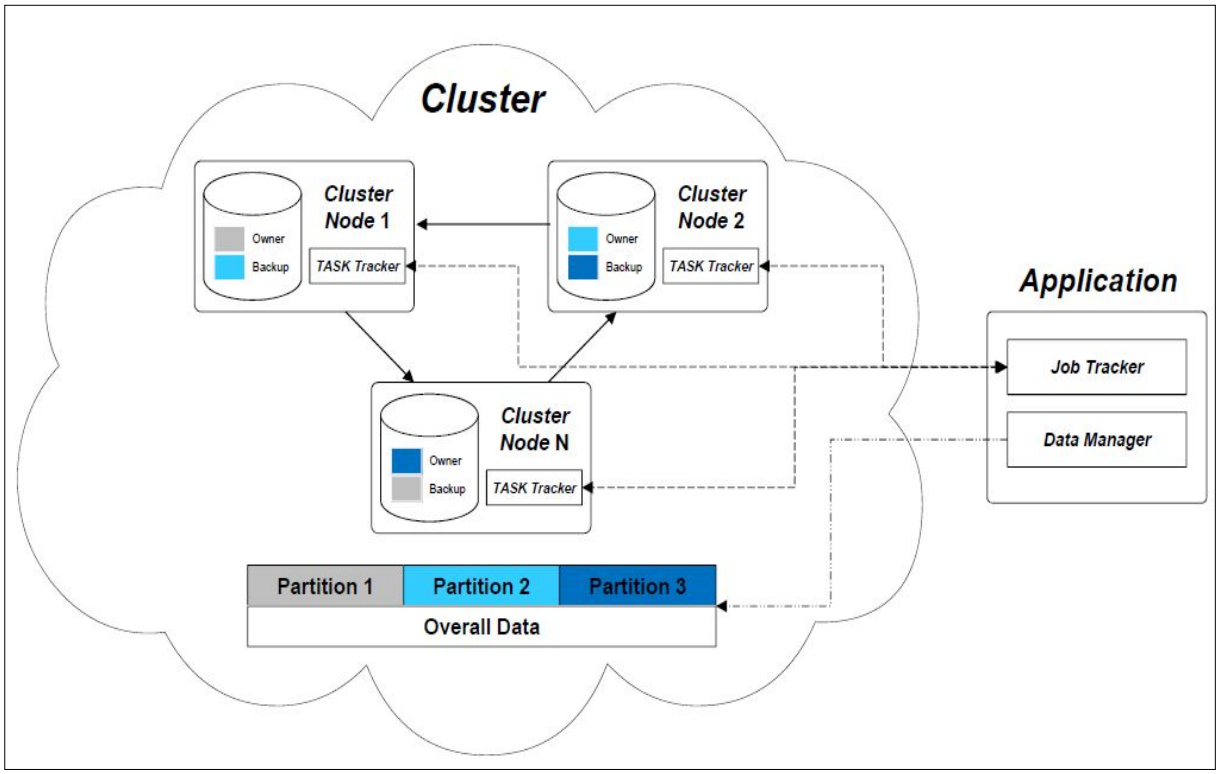


그림 8. 개선된 IMDG 기반 염기서열 클러스터링 프로그램 아키텍처 모식도

다. 클러스터는 분산 환경에서의 n개 이상의 클러스터 노드(*Cluster Node*)가 합쳐진 것임. 각각의 클러스터 노드는 서열 클러스터링을 위한 연산작업을 수행하며, 동시에 데이터를 저장하는 역할을 함. 이때 각 노드의 데이터는 Owner와 Backup영역으로 전체 데이터를 서로 나눠져 저장되어 있으며, 하나의 노드가 가동 중단되더라도 다른 노드에 데이터가 저장되기 때문에, 데이터의 안정성이 보장되도록 하였음.

라. 태스크트래커(*Task tracker*)는 특정 작업에 대한 논리적인 작업단위로서 수행될 연산의 정의, 사용 스레드의 개수, 데이터의 범위, 결과의 저장위치 등에 대한 정보가 담겨 있음. 태스크트래커는 어플리케이션에 의해 분산환경내의 각 노드에 인스턴스로서 할당되며 해당 노드의 가용 개수만큼의 멀티 threads를 발생시켜 정의된 task를 병렬로 빠르게 처리하도록 함.

마. 어플리케이션은 클러스터와의 통신을 유지하며 사용자로 부터 특정 작업 요청을 받아 이를 처리하는 역할을 수행. 어플리케이션은 잡트래커(*Job tracker*)와 데이터매니져(*Data manager*)로 나뉨. 잡트래커는 각 노드에 태스크트래커를 생성, 배분하는 역할을 수행하며 가 노드에 할당된 태스크트래커 의 수행 결과 및 상태를 지속적으로 체크, 태스크 수행 중 장애가 발생한 노드의 태스크트레커를 다른 노드에 자동으로 할당하도록 하여 고가용성을 보장하도록 구현됨.

바. 데이터매니져를 구현하여 메모리 누수 및 부족 문제(out of memory)가 발생하지 않도록 처리도중 산출된 더 이상 불필요한 중간데이터를 삭제하는 역할을 수행 하도록 함

## 제 3-2절: Whole 메타전사체 shotgun 염기서열 분석 파이프라인 구

축

1. 자동화된 메타지놈 염기서열 어셈블리(assembly) 파이프라인 구축

가. 일반적으로 일루미나 기반 시퀀서는 적은 비용으로 많은 데이터의 가능하기 때문에 높은 커버러지가 담보되야하는 홀 지놈 샷건 시퀀싱을 통한 메타지놈 연구에 많이 사용됨. 그러나 일루미나 기반 데이터는 전체 유전자 영역을 커버하지 못하는 short read 이기 때문에 정확하게 데이터를 분석하기 위해서는 어셈블리(assembly)과정이 필수적으로 동반되어야 함.

나. 현재 short read 기반의 메타지놈 서열 어셈블리를 위해 Meta-IDBA (Peng et al. 2011), MetaVelvet (Namiki et al. 2012), IDBA-UD (Peng et al. 2012), Ray Meta (Boisvert et al. 2012) 등의 여러 가지 도구들이 개발되었으나, 메타지놈 연구에 적용하기에 정확도가 떨어지는 문제가 있음. 따라서 본 연구팀은 이제 까지 소개된 모든 어셈블리 도구들의 장단점을 분석하여, 가장 정확도가 높다고 판단되는 최근 소개된 메타지놈 어셈블리 도구인 Meta-velvetSL (Sato et al. 2015)을 기반으로 하는 어셈블리 파이프라인을 구축함. 본 파이프라인은 Meta-velvetSL을 기반으로 하였기 높은 정확도를 갖는 contig의 생산이 가능.

다. Meta-velvetSL은 여러 외부 도구가 필요하고 여러 가지 단계를 거쳐야하는 복잡한 구동 과정으로 인해 사용이 매우 어려운 단점이 있음. 이에 본 연구팀은 자동화된 어셈블리 파이프라인을 구축하여 사용자가 편리하게 구동이 가능하도록 함(그림 9).

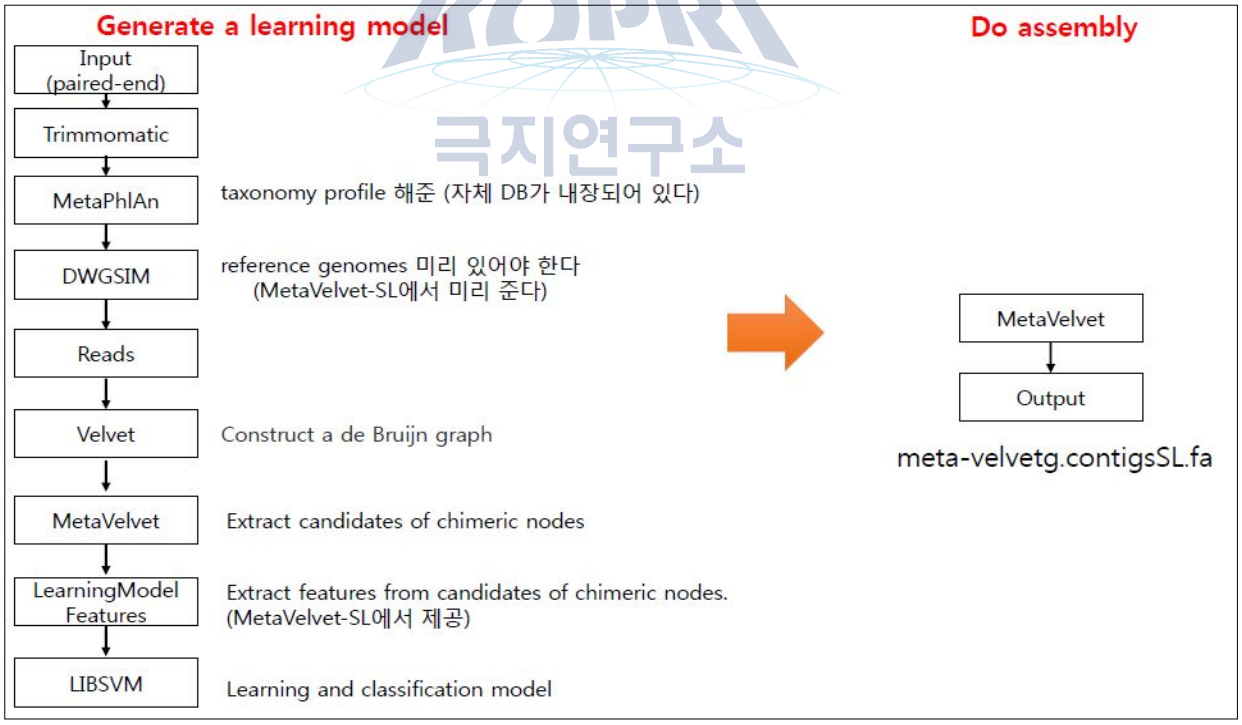


그림 9. 메타지놈 염기서열 어셈블리 파이프라인 워크플로우

라. 본 파이프라인은 쉘 스크립트 및 리눅스에서 사용하는 명령어를 조합하여 구현하였기 때문에 유지보수가 쉽고 즉각적인 추가기능의 탑재가 가능 (그림 10).

```
#!/bin/bash

#

#bash    meta_new_kdlsh.sh    test1    ERR011104_1.fastq.gz    ERR011104_2.fastq.gz    $(pwd)    70
/home/ofang/Database

#bash                                meta_new_kdlsh.sh                                testseven
/home/ofang/shotgun/MetaVelvetSL_Pipeline/sra_data/seven_1.fastq.gz
/home/ofang/shotgun/MetaVelvetSL_Pipeline/sra_data/seven_2.fastq.gz          $(pwd)          100
/home/ofang/Database &> meta_new_kdlsh.sh.log

#

if [ $# -ne 6 ]

  then

    echo "USAGE Example: bash meta_new_kdlsh.sh test1 ERR011104_1.fastq.gz ERR011104_2.fastq.gz
#$(pwd) 70 /home/ofang/Database"

    exit 1

fi

#

project=$1
```

그림 10. 염기서열 어셈블리 파이프라인 쉘 스크립트

마. 본 파이프라인은 Paried-end fastaq서열을 입력값으로 받아들이며 추가적인 fasta 파일 형태로의 변환과정 없이 바로 실행이 가능하도록 구현.

바. 일루미나 기반의 short read 서열 데이터에는 여러 가지 오류가 있거나 퀄리티가 낮고, 길이가 짧은 서열이 섞여 있을 수 있음. 이러한 데이터는 추후 어셈블리과정에서 잘못된 결과를 도출할 수 있기 때문에 제거되어야 함. 본 파이프라인에서는 요즘 가장 널리 사용되며, 정확하다고 알려진 Trimmomatic (Bolger et al. 2014) 을 탑재하여 서열 전처리 과정을 통해 정확한 어셈블리가 가능하도록 함.

사. 본 파이프라인은 Meta-velvetSL을 사용하였기 때문에 chimera filtering이 가능함.

아. 실제 메타지놈 샷건 데이터를 가지고 본 파이프라인의 구동 테스트를 수행. 데이터 셋은 NCBI SRA로부터 다운 받은 일루미나 HiSeq 2000 기반 paired-end 데이터 (SRR942940, SRR943323) 40G 용량의 압축된 fastaq 형태 임. 실험 환경은 AMD Opteron(TM) Processor 6274 2.2 MHz 64 core CPU와 512GB 메모리를 탑재한 리눅스 서버에서 수행하였으며, 34 시간 만에 성공적으로 어셈블리가 가능함을 보임.

2. 메타지놈 유전자 예측 및 기능 annotation 위한 분석 파이프라인

가. 서열 어셈블리 과정을 통해 생성된 contig 서열을 가지고 유전자 예측 및 annotation을 위한 분석 파이프라인을 구축함. 본 파이프라인은 어셈블리 파이프라인과 같이 쉘 스크립트 및 리눅스에서 사용하는 명령어를 조합하여 구현하였기 때문에 유지보수가 쉽고 추가기능의 탑재가 용이.

나. 본 파이프라인은 서열 어셈블리 과정 이후, 먼저 contig 서열에서의 유전자 예측과정을 거쳐 유전자의 위치와 구조가 파악되었으면 그 유전자의 서열정보를 통해 유전자의 기능을 유추 하는 과정을 진행함 (그림 11).

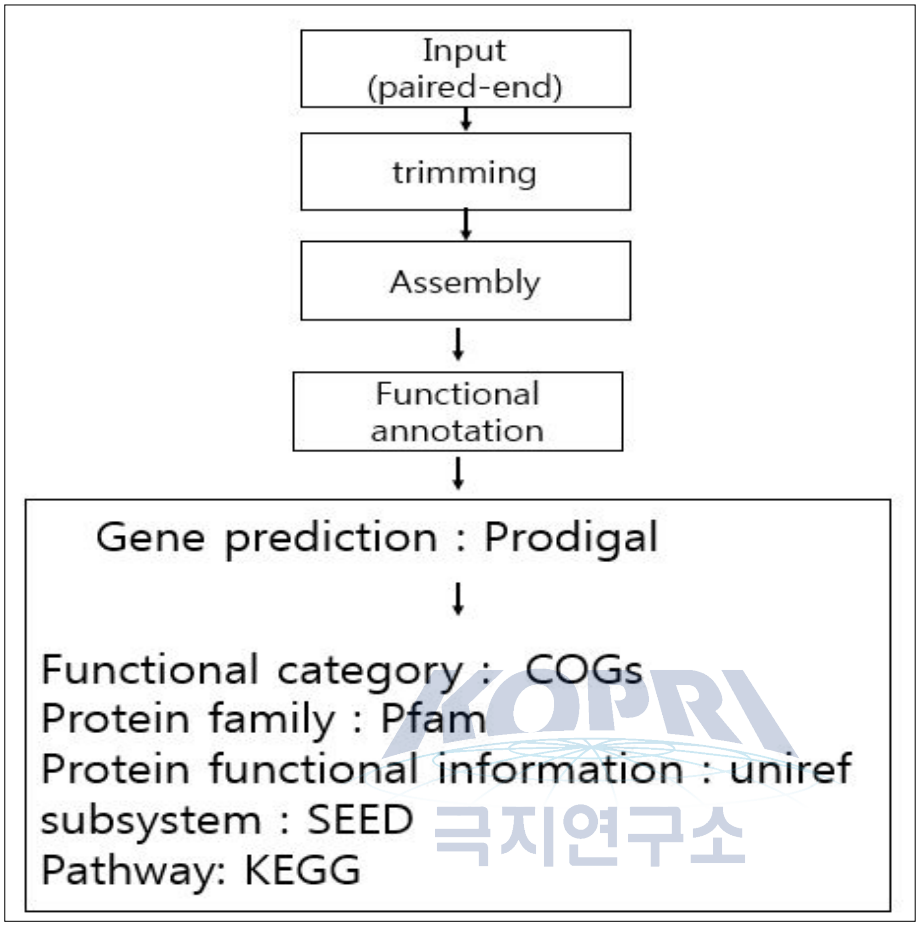


그림 11. 메타지놈 유전자 예측 및 기능 annotation 위한 분석 파이프라인 워크플로우

다. 이 과정은 서열의 상동성을 기반으로 수행되며, blast를 사용하여 유전자의 기능을 유추하게 됨. rpsblast와 pfam 데이터베이스를 사용 단백질의 기능 유추, blastp와 UniProt 데이터베이스를 사용하여 단백질의 기능 유추 함. 단백질 기능에 대한 계층별로 기능을 유추하기 위해 blastp와 subsystem 데이터베이스를 사용 함. 해당 단백질 서열의 Orthologous 그룹 정보를 통한 기능 유추를 위해 blastp와 COGs 데이터베이스를 사용 함. 환경 유전체 단백질 서열에 대한 메타볼릭 패스웨이 분석을 진행하기 위해 blastp와 Kegg데이터베이스에를 사용함. 본 연구팀은 이미 kegg데이터베이스에 대한 라이센스를 구비한 상태임. 본 파이프라인은 이러한 다양한 annotation과정을 통해 사용자에게 쉽고 편리하게, 정확하고 다양한 정보를 제공이 가능함.

라. 본 파이프라인은 단백질 서열에서 signal peptide cleavage sites를 유추하기 위해 signalp (Petersen et al. 2011)를 사용하는데, signalp는 10000이상의 서열은 처리하지 못하는 문제점이 있음. 이에 본 연구팀은 처리해야 할 서열을 10000씩 분할해서 처리하도록 구현하여 이 문제를 해결함.

마. Annotation 수행 후 나온 모든 분석결과들은 자체 개발되어 파이프라인에 탑재된 Java

기반 프로그램에 의해 tab 분할 형식의 파일 형태로 요약되어 사용자들이 결과를 파악하기 쉽게 함(그림 12). 출력되는 정보는 Contig_ID, Start, End, Strand, Uniref100, CogNo, Cog_Category, Pfam, Gene_ID, Kegg Des, Kegg pathway 순으로 출력되며 Kegg pathway 정보는 하나의 Gene_ID에 여러개의 pathway id 가 존재할 수 있기 때문에 대괄호([]로 해당 pathway id를 묶어서 표현한다(표 2).


그림 12. Annotation 분석결과

표 2. Annotation 수행 후 나온 모든 분석결과내용 요약

| 속성 | 예시 |
|---|---|
| Contig_ID | NODE_49_length_590533_cov_18.955446_38 |
| Start | 41841 |
| End | 42929 |
| Strand | - |
| Uniref100 | 2-aminoethylphosphonate--pyruvate transaminase |
| CogNo | COG0075 |
| Cog_Category | E |
| Pfam | pfam00266 Aminotran_5 Aminotransferase class-V |
| Gene_ID | bvu:BVU_3006 |
| Kegg Des | 2-aminoethylphosphonate--pyruvate transaminase |
| Kegg pathway | [bvu00440, bvu01100, bvu01120] |

# 제 4 장 연구개발목표 달성도 및 대외기여도

## 제 4-1절: 2015 2차년도

| 연구목표 | 연구내용 | 달성도 | 대외기여도 |
|---|---|---|---|
| Whole genome shotgun 메타전사체 데이터 분석 프로그램 개발 | In-memory 기반 염기서열 클러스터링 클라우드 환경 구축 고도화 | 100% | - In-memory 기반 염기서열 클러스터링 클라우드 환경에서 구동 되도록 고도화되었는지 완료.<br>- 기존에 개발된 JAVA 기반 In-Memory Data Grid (IMDG) 분산 처리 기능의 고가용성을 개선 완료.<br>- 클라우드 환경 구동 및 쉬운 사용성을 위한 사용자 편의 기능 개선 완료. |
| | Whole 메타전사체 shotgun 염기서열 분석 파이프라인 구축 | 100% | - 메타지놈 염기서열 assembly 파이프라인 구축완료.<br>- 메타지놈 유전자 예측 및 기능 annotation 을 위한 분석 파이프라인 구축완료. |

# 제 5 장 연구개발결과의 활용계획

## 제 5-1절: 추가연구의 필요성

1. Whole 메타전사체 shotgun 염기서열에 대해 해당 환경에서의 미생물의 역할에 대해 더 깊게 이해하기 위해서는 functional annotation과 더불어 메타지놈 내 특이 유전자 클러스터링 분석 및 빈도 비교 분석이 필요함. 본 연구팀은 올해 구축된 파이프라인을 바탕으로 특이 유전자 클러스터링 분석 및 빈도 비교 분석을 위한 다양한 생물적보학적 연구를 수행할 예정임.

   가. RNA-Seq 데이터에서 사용하는 FPKM (fragment per kilobase of transcript per million fragments)의 개념을 차용하여, 유전체 내 1 copy만을 지니는 유전자들과 관심있는 특이 유전자들의 reference 유전자 염기서열의 단위 길이당 match되는 read 수를 파악하여 해당 시료 내 특이 유전자를 지니는 미생물의 빈도수를 계산할 수 있도록 통계방법 구현이 필요함.

   나. 임의로 단편화되어 align이 되지 않는 메타지놈 reads를 clustering하기 위하여 해당되는 각 read의 염기서열에 표준화된 정보값을 부여하고 이를 character value로 인지하여 clustering을 수행할 수 있는 체계를 구현함. 이로부터 시료 내 유전자 그룹별 빈도를 파악할 수 있는 자동화 분석 시스템 개발함.

   다. 메타지놈 서열로부터 유전자의 빈도수 계산을 위한 프로그램을 개발하고, 일련의 과정들을 서로 연동하여 standalone 수준에서 분석 파이프라인을 구축함.

2. 본 과제 수행을 통해 Whole 메타전사체 shotgun 염기서열 분석 파이프라인을 구축하였고 일루미나 기반의 메타지놈 서열에 대해 편리하게 functional annotation이 가능하게 됨. 본 연구팀은 이에 덧붙여 주어진 환경상에서의 미생물 역할을 더 깊게 이해하기 위해 metabolic pathway 관련 유전자 빈도 비교 분석 파이프라인 구축할 예정임.

   가. 확장형 KEGG database의 염기서열과 메타지놈 염기서열 함께 CLUSTOM 알고리즘을 이용하여 clustering한 후, 각 cluster에서 KEGG ID를 추출하고 rpoB, recA 유전자에 match된 수 및 각 reference 유전자 그룹의 size로 정규화하여 메타지놈 내 각 metabolism에 관여하는 유전자들의 빈도수 계산 방법을 구축함.

   나. 구축된 방법을 이용하여 KEGG ID에 따른 빈도수를 자동으로 계산하는 프로그램을 구현하고 table 형식으로 도출할 수 있도록 설계함.

   다. Hypergeometric distribution, t-test등의 통계적 방법을 이용하여 샘플별 enriched된 세부 metabolic pathway 정보 표현

   라. KEGG ID 및 EC number를 선택하여 해당하는 metagenome reads를 다운로드할 수 있도록 설계함

   마. 분석 시료의 메타데이터(시료 채취 날짜, 온도, 위도 및 경도, biogeochemical data 등)를 연동하여 보여줄 수 있도록 설계함

## 제 5-2절: 기업화 추진방안

1. 최근 NGS 기술의 발전으로 생물학 분야에서 대량의 염기서열 데이터가 생산되고 있음. 이들 데이터는 양적인 측면에서 단일 연구실에서 분석하기가 힘들어 회사 수준에서의 상용서비스가 필요한 실정임. 본 과제에서 개발된 프로그램 및 데이터베이스는 추후 관련 분야 연구자들이 상용서비스로 이용될 가능성이 있음. 하지만, 현 단계에서 구체적인 기업화 추진방안을 논의하는 것은 시기상조로 판단됨.

# 제 6 장 참고문헌


Scholloss PD, et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537-7541.

Kuczynski, Justin, et al. "Using QIIME to analyze 16S rRNA gene sequences from microbial communities." Current protocols in microbiology (2012): 1E-5.

Glass, Elizabeth M., et al. "Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes." Cold Spring Harbor Protocols 2010.1 (2010): pdb-prot5368.

Markowitz, Victor M., et al. "IMG/M: a data management and analysis system for metagenomes." Nucleic acids research 36.suppl 1 (2008): D534-D538.

Huson, Daniel H., et al. "Integrative analysis of environmental sequences using MEGAN4." Genome research 21.9 (2011): 1552-1560.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.

Rodrigues, J. F. M., & von Mering, C. 2013. HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. Bioinformatics, btt657.

Hwang K, Oh J, Kim T, Kim BK, Yu DS, Hou BK, Caetano-Anolles G, Hong SG, Kim KM. 2013. CLUSTOM: a novel method for clustering 16S rRNA next generation sequences by overlap minimization. PLOS ONE 8:e62623.

Tatusov, Roman L., et al. "The COG database: an updated version includes eukaryotes." BMC bioinformatics 4.1 (2003): 41.

Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research 28.1 (2000): 27-30.

Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource." Nucleic acids research 32.suppl 1 (2004): D258-D261.

Meyer, Folker, Ross Overbeek, and Alex Rodriguez. "FIGfams: yet another set of protein families." Nucleic acids research 37.20 (2009): 6643-6654.

Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. (2009) The NIH human microbiome project. Genome research 19: 2317-2323.

Sato, Kengo, and Yasubumi Sakakibara. "MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised

learning." DNA Research 22.1 (2015): 69-77.

Peng Y., Leung H.C.M., Yiu S.M., et al. Meta-IDBA: a de novo assembler for metagenomic data. Bioinformatics 2011;27:i94-101.

Namiki T., Hachiya T., Tanaka H., Sakakibara Y. Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40:e155.

Peng Y., Leung H.C.M., Yiu S.M., et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 2012;28:1420-8.

Boisvert S., Raymond F., Godzaridis E., et al. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 2012;13:r22.

Sato, K., & Sakakibara, Y. (2015). MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Research, 22(1), 69-77.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, btu170.

Petersen, Thomas Nordahl, et al. "SignalP 4.0: discriminating signal peptides from transmembrane regions." Nature methods 8.10 (2011): 785-786.

주 의

1. 이 보고서는 극지연구소 위탁과제 연구결과 보고서입니다.
2. 이 보고서 내용을 발표할 때에는 반드시 극지연구소에서 위탁연구과제로 수행한 연구결과임을 밝혀야 합니다.
3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 안 됩니다.