***Idotea baltica* Assembly Clustering**

CD-HIT (v4.6.1)

**RUNNING**

Load biokit

module load biokit

Create batch job

#!/bin/bash -l

# created: Dec 10, 2015 11:21 AM

# author: ketaya

#SBATCH -J cd-hit

#SBATCH -n 1

#SBATCH -o out\_%j.txt

#SBATCH -e err\_%j.txt

#SBATCH --cpus-per-task=16

#SBATCH -t 48:00:00

#SBATCH --mem=16000

#SBATCH --mail-type=END

#SBATCH --mail-user=keith.t.yamada@utu.fi

module load biokit

cd-hit-est -i trinity.all.fasta -o trinity.all.nr95.fasta -c 0.95 \

-n 10 -d 0 -M 16000 -T 16

\*\*\* No need for batch job file. It only takes 5min.

Run

================================================================

Program: CD-HIT, V4.6 (+OpenMP), Apr 18 2013, 12:25:50

Command: cd-hit-est -i trinity.all.fasta -o

trinity.all.nr95.fasta -c 0.95 -n 10 -d 0 -M 16000 -T

16

Started: Thu Jan 7 20:40:22 2016

================================================================

Output

----------------------------------------------------------------

total seq: 309529

longest and shortest : 21569 and 201

Total letters: 203,175,509

Sequences have been sorted

Approximated minimal memory consumption:

Sequence : 243M

Buffer : 16 X 20M = 325M

Table : 2 X 21M = 43M

Miscellaneous : 7M

Total : 620M

Table limit with the given memory limit:

Max number of representatives: 4000000

Max number of word counting entries: 1922400508

# comparing sequences from 0 to 17196

.......... 10000 finished 4129 clusters

.....---------- new table with 7001 representatives

# comparing sequences from 17196 to 33436

.......... 20000 finished 8149 clusters

---------- 8168 remaining sequences to the next cycle

---------- new table with 3228 representatives

# comparing sequences from 25268 to 41060

.......... 30000 finished 12097 clusters

---------- 10865 remaining sequences to the next cycle

---------- new table with 1942 representatives

...

309529 finished 115917 clusters

Apprixmated maximum memory consumption: 836M

writing new database

writing clustering information

program completed !

Total CPU time 1079.27

\*\*\* ~ 5 min

**RESULTS**

Some headers are duplicated and need to be renamed for transrate to work

python rename\_fasta.py trinity.all.nr95.fasta \_ trinity.all.nr95.renamed.fasta

\*\* ~

\*\*