Verification of the Ultrafast Clustering Algorithm CD-HIT

Gregory Sprenger

Depart of Biotechnology, Johns Hopkins University

ABSTRACT

Being able to replicate a study is an essential factor of published research. This study aims to replicate the study "Ultrafast clustering algorithms for metagenomic sequence analysis" to determine if it is replicable and to verify the results presented. It was found that the study was fairly replicable, and the results were verified for the data that was attainable.

Keywords: cd-hit; cd-hit package; cd-hit-est; clustering; OTU; operational taxonomic units; metagenomics

INTRODUCTION

Metagenomics is a rapidly advancing genomic approach that aims to study the microorganisms of an environmental sample. In part to the rapid advancements of metagenomics is due to the advancements in DNA sequencing. Metagenomic samples can have thousands of microorganisms present, and therefore next-generation sequencing (NGS) technologies, such as Illumina, have greatly increased not only the turnaround times but also increased the sensitivity to be able to detect low-frequency variants in samples (Illumina, 2020). This leads to the bottleneck of metagenomics, where the complexity and quantity of the data from NGS technologies creates a challenge for data analyses. Such challenges are variations on how the samples are extracted and handled, differing genome sizes in the sample, and sequencing errors or variations, which all affect the ability for analytic tools to identify genes in the sample (Morgan, Darling, & Eisen, 2010). To combat these issues, several algorithms were developed to aide in the removal of errors and variability in sequences. One such algorithm is called cluster analysis: a machine learning method that that aims to discover groupings in the data. CD-HIT is a program used for clustering and sequence comparisons of large datasets to decrease redundancy and improve the performance of other analytical tools (Fu, Niu, Zhu, Wu, & Li, 2012).

One of the purposes of published research is the ability to reproduce and verify the information presented. This is to ensure that the procedures listed can be successfully reproduced and generate the same or similar results as the original. The aim of this study is to determine the ability to reproduce and verify the results of CD-HIT from the article "Ultrafast clustering algorithms for metagenomic sequence analysis" (Li, Fu, Niu, Wu, & Wooley, 2012).

METHODS

Analyses were conducted on Google cloud VM instances and had Ubuntu 18.04 LTS as the operating system. The instances had either 4 CPUs and 32 GB memory or 8 CPUs and 64 GB memory. Comparable software versions of programs that were used in the original study were obtained and used. If no version number were specified in the original study, then the most up to date version was used. For this study, CD-HIT v4.5.7 and CD-HIT-OUT-454 v0.0.2 were used and were obtained from Google's archive. Data files were pulled from direct links from the original study, the European Read Archive, or the Short Read archive. Appendix A lists the information needed to obtain data files that were used in this study.

This study aims to replicate the date processing of the datasets in table one through four of the original study using the same parameters listed (Li, Fu, Niu, Wu, & Wooley, 2012). If no parameters were listed, the programs default parameters were used. Time was calculated for all datasets that were analyzed on the four CPU instances by dividing the time in minutes by four to get an approximate time if the data were ran on one thread.

RESULTS

Due to docker containers not allowing the compilation of the multithreaded versions of CD-HIT, the datasets were processed one at a time on the VM instances with each program to avoid variations in the analysis. The analyses of the datasets with various programs are not intended for performance comparisons but are used to show the differences in the speed and clustering capabilities of these programs (Li, Fu, Niu, Wu, & Wooley, 2012).

Table 1: Speed of clustering algorithms on datasets

Dataset	**Program (percent identity)**	**Original Time (minutes)**	**Tested Time (minutes)**	**Original Clusters**	**Tested Clusters**
NCBI NR	CD-HIT (90)	1405	825	7036029	10565347
	CD-HIT (70)	962	705	4933074	7178871
Swissprot proteins	CD-HIT (90)	3.7	4	298617	309315
	Uclust (90)	17.3	5	301076	308340
	CD-HIT (70)	4.6	5	190695	198739
	Uclust (70)	7.6	5	192847	197099
Illumina SRR061270	CD-HIT (95)	56.8	168	956734	3918401
	Uclust (95)[a]	164.6	569	958887	7244469
	CD-HIT (90)[e]	347.5	[redacted]	751581	[redacted]
	Uclust (90)[b]	227.5	699	734981	6825702
	CD-HIT (90)[c]	23.5	[redacted]	750276	[redacted]
	SEED (default)[d]	7.9	[redacted]	1056109	[redacted]
Human body 16S rRNA	CD-HIT (97)	47.9	12	24842	22347
	Uclust (97)	4.3	3	29586	25331
	DNACLUST (97)	15.3	9	31151	27872

[a,b]Illumina datafile was split into five pieces due to constraints of the free Uclust. Time and clusters are the combination of all five datafiles. [L]atest version of CD-HIT is 4.8 and no version 5 beta was found. [d]Length of sequences were too large for SEED to cluster. All analyses were conducted on 4 core 32GB memory VM instances. [e]CD-HIT at 90% identity was not completed at time of publishing.

Data	True OTUs	Predicted OTUs		Time (seconds)	
		Original	Tested	Original	Tested
Divergent	23	26	26	11	4
Artificial	33	32	34	13	4
Even1	53	71	75	8	2
Even2	53	57	59	7	2
Even3	52	60	63	7	2
Uneven1	49	56	55	5	2
Uneven2	41	45	46	7	2
Uneven3	38	42	40	7	2
Titanium	69	69	94	7	4
Human gut		317	138	37	17
Human body		238	239	295	151

Table 2: Comparison of speed and accuracy of the identification of OTUS

All analyses were conducted on 4 core 32 GB memory VM instances.

All of the datasets that were obtained for this study were larger than that of the original study. Interestingly, each of the programs were able to identify similar or more clusters in similar or faster times than the original study. This may be due to the greedy incremental approaches that these programs use. SEED was not able to be used due to its sequence length cap of 100 base pairs and the Illumina dataset having an average sequence length of 152 base pairs.

The free 32-bit version of Uclust only allows the use of 4GB memory, and therefore was not enough memory to complete the analysis of the Illumina reads. The 64-bit version of Uclust allows the allocation of all of a computer's memory, but this software was not available for use. Therefore, to get an approximation of the total runtime and clusters from Uclust, the dataset was converted into FASTA format and the total number of lines were computed using bash commands. This allowed the dataset to be split into five files, each containing 10 million lines (or 5 million sequences). Each file was ran separately and concatenated to give the number of clusters and runtime in Table 1.

The accuracy and speed of CD-HIT-OTU were compared with the results of the original study (Table 2). CD-HIT-OTU is a combination of the tools CD-HIT-DUP and CD-HIT-EST. CD-HIT-DUP is used to cluster the reads by identifying duplicate reads on Illumina datasets. The data used in the original study are 454 reads and therefore CD-HIT-DUP may not be as efficient as CD-HIT-454. Though the CD-HIT package does not allow changing of algorithms in the CD-HIT-OTU script. After the duplicate reads are identified and clustered, CD-HIT-EST is used to cluster the nucleotide sequences into OTUs (Li, Fu, Niu, Wu, & Wooley, 2012). The data shows that they number of predicted OTUs of the datasets are similar, besides the results of the Titanium dataset. Though, the previous study does show AmpliconNoise and Denoiser having much more comparable results as this study on the Titanium dataset (Li, Fu, Niu, Wu, & Wooley, 2012). Many of the reference sequence datasets were unobtainable and therefore the true OTUs were not calculated, as well as being able to verify the results of the AmpliconNoise and Denoiser analyses. Instead, Qiime1 was used to calculate OTUs after

Table 3: Comparison of clustering and file reduction of reference databases using CD-HIT

Data	**Original Percent Reduction**	**Tested Percent Reduction**
NCBI NR	58	44
16S (Silva + Greengene)	28	85
NCBI Microbial Genomes	38	
NCBI Virus Sequences	28	
Human body		99
Human gut		99
Human body and gut concatenated		99
NCBI Influenza		99

All analyses were computed on 8 core 64 GB memory VM instances.

clustering with CD-HIT and can be found in Appendix B. OTUs were also conducted on pooled human metagenomic samples in the original study and therefore replicated and presented in Table 2. The human body dataset was successfully reproduced, but the human gut dataset showed variation. This may be due to how the 815 datafiles were obtained from the European Read Archive and then concatenated together.

Various CD-HIT programs in the CD-HIT package were tested on different reference genomes to determine the time to cluster the genomes, as well as determine the percent reduction of files. Due to the differences in the file size and therefore the number of sequences in the original study to the data obtained for this replication study, only the percent reduction of the files are compared. A data table with all information on the reference database clustering can be found in Appendix B.

Interestingly, the NBI NR dataset used in this study was roughly twice as large as the dataset in the original study but was still able to decrease its overall size by almost half after clustering with CD-HIT. Due to virus sequence database being unattainable for this study, the NCBI influenza dataset was used due to it being a virus as well as having a similar file size of the virus database. The NCBI microbial data sets were obtained but were not able to be processed by the CD-HIT package due to the long sequence lengths that resulted in segmentation faults in the CD-HIT programs. Therefore, the human body and gut data sets from the OTU identification were used. The file size of the 16S Silva and Greengene data were approximately 160 MB less than the original study, though the percent reduction was almost 86% which is almost three times that of the original study.

CONCLUSIONS

One of the tenants of published research is to be able to replicate the study to verify the results presented. In bioinformatics, this means being able to obtain data files and programs used to be able to rerun the analyses. The data files from “Ultrafast clustering algorithms for metagenomic sequence analysis” used links to many data files that are no longer available. This made it hard to obtain the same data files and therefore similar data files were obtained. This has ultimately led to skewed results on many of the analyses that were conducted in this study. The original study provided all of the parameters to replicate its study but failed to mention which program in the CD-HIT package was used. This gave a false sense that each data set was analyzed using the default CD-HIT program. Therefore, during this study, the proper clustering program had to be referenced for each type of data set.

Further studies need to be conducted to examine the side-by-side performance of CD-HIT and other clustering algorithms. Such clustering algorithms that would be of interest would be VSEARCH which is based on USEARCH and is free for public use in either 32- or 64-bit versions. Uclust, the clustering part of USEARCH was used in Table 1 and would be interesting to see how they both differ from CD-HIT. GeFaST, an OTU identifier that uses Swarm's clustering approach (Muller & Nebel, 2018). GeFaST may be a better tool than CD-HIT due to it being focused on 16S rRNA sequences, whereas CD-HIT focuses on removing duplicates and then using a nucleotide clustering algorithm to identify OTUs. Modern denoiser algorithms should also be examined in a side-by-side performance of CD-HIT, such as DADA2, Deblur, and UNOISE3. A recent study shows that the various OTU identifiers reported different results and therefore bioinformatic tools should be carefully selected when attempting to discover rare variants (Nearing, Douglas, Comeau, & Langille, 2018).

REFERENCES

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., & Knight, R. (2009, December 18). Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science, 326*(5960), 1694-1697. doi:10.1126/science.1177486

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012, December 1). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics, 28*(23), 3150-3152. doi:doi.org/10.1093/bioinformatics/bts565

Illumina. (2020). *NGS vs. Sanger Sequencing*. Retrieved December 13, 2020, from Illumina: https://www.illumina.com/science/technology/next-generation-sequencing/ngs-vs-sanger-sequencing.html

Koxlowski, L. P. (2018, May 14). *Biological databases*. Retrieved December 10, 2020, from IIMCB, Laboratory of Bioinformatics and Protein Engineering: ftp://genesilico.pl/lukaskoz/biological_databases/

Li, W., Fu, L., Niu, B., Wu, S., & Wooley, J. (2012, July 6). Ultrafast clustering algorithm for metagenomic sequence analysis. *Briefings in Bioinformatics*, 656-668. doi:10.1093/bib/bbs035

Morgan, J. L., Darling, A. E., & Eisen, J. A. (2010, April 16). Metagenomic Sequencing of an In Vitro-Simulated Microbial Community. *PLOS One, 5*(4). doi:https://doi.org/10.1371/journal.pone.0010209

Muller, R., & Nebel, M. E. (2018, September 12). GeFaST: An improved method for OTU assignment by generalising Swarm's fastidious clustering approach. *BMC Bioinformatics*. doi:10.1186/s12859-018-2349-1

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. (2018, August 8). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*. doi:10.7717/peerj.5364

Weizhong Lab. (2020). *CD-HIT-OTU Download*. Retrieved from Weizhong Lab: http://weizhong-lab.ucsd.edu/cd-hit-otu/download.php

APPENDIX A

Datasets

Datasets were pulled from various resources to obtain datafiles as close to the time of publication of the original study. To determine the speed of clustering programs the NR and Swissprot datasets were downloaded from a third party archive database at ftp://genesilico.pl/lukaskoz/biological_databases/ due to NCBI not archiving old databases (Koxlowski, 2018). The Illumina reads from SRR061270 were pulled using the SRA Toolkit. The 16S human body and human gut rRNA reads are from another study and were pulled from the European Read Archive with the accession numbers ERA000159 and PRJNA32089, respectively. After pulling each data file, the data in each file were concatenated.

The datasets used for the determination of the accuracy and speed of OTUs were pulled from the CD-HIT-OTU download page: http://weizhong-lab.ucsd.edu/cd-hit-otu/download.php (Weizhong Lab, 2020).

For the evaluation of clustering on reference databases, the NCBI influenza dataset, influenza.fna, were pulled from https://ftp.ncbi.nih.gov/genomes/INFLUENZA/ and the NCBI bacteria dataset, all.fna.tar.gz, were pulled from https://ftp.ncbi.nih.gov/genomes/archive/old_refseq/Bacteria/. The Silva and Greengene datasets were pulled from https://www.arb-silva.de/no_cache/download/archive/vectordb/ (Silva_vector_db_release_111.fasta) and http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/ (current_GREENGENES_gg16S_unaligned.fasta.gz). The influenza.fna dataset was obtained from https://ftp.ncbi.nih.gov/genomes/INFLUENZA/.

Tools

CD-HIT was obtained from Google's archive: https://code.google.com/archive/p/cdhit/downloads. CD-HIT required editing of line 95 of cdhit-common.h so that it becomes "this->push_back(item);". CD-HIT-OTU also required editing of a file, minArray.hxx, so that "#include<stdio.h>" and "#include<string.h>" were incorporated at the top of the file. Once these edits were conducted, the programs were able to be successfully compiled and used.

DNACLUST release 3 was obtained from https://sourceforge.net/projects/dnaclust/files/ and compiled on the VM instance.

SEED was obtained from https://github.com/baoe/SEED and compiled on the VM instance. SEED does not contain version numbers.

SraToolkit version 2.10.8-ubuntu64 was pulled from the Short Read Archive to obtain Illumina and human gut data files.

USEARCH 5.2.32 was used and can be obtained from https://www.drive5.com/usearch/download.html.

Qiime was ran via docker container and can be pulled from using the following command: docker pull bwawrik/qiime:v3.

APPENDIX B

Table 1A: Analysis of clustering speed on data sets

Data	Number of Sequences	File Size	Program and Parameters	Time (minutes)	Clusters
NCBI NR	18599335	9.9 GB	cd-hit '-n 5 -M 0 -c 0.9 -T 0'	825.4	10565347
			cd-hit '-n 5 -M 0 -c 0.7 -T 0'	705.4	7178871
Swissprot proteins	451845	231 MB	cd-hit '-n 5 -M 0 -c 0.9 -T 0'	4	309315
			uclust '-id 0.9'	5.2	308340
			cd-hit '-n 5 -M 0 -c 0.7 -T 0'	4.7	198739
			uclust '-id 0.7'	4.7	197099
Illumina SRR061270	20945329	8 GB	cd-hit '-n 10 -M 0 -c 0.95 -T 0'	168.3	3918401
			uclust '-id 0.95'[a]	569	7244469
			cd-hit '-n 10 -M 0 -c 0.9 -T 0'[e]	■	■
			uclust '-id 0.9'[b]	699.2	6825702
			cd-hit v5 beta '-c 0.9'[c]	■	■
			SEED (default)[d]	■	■
16S rRNA reads	1071335	317 MB	cd-hit '-n 10 -M 0 -c 0.97 -T 0'	11.9	22347
			uclust '-id 0.97'	2.6	25331
			DNACLUST '-s 0.97'	9	27872

All analyses were computed on 4 core 32 GB memory VM instance. [a, b]Uclust data is the combination of SRR061270 being split into five files containing 5 million sequences each. Data can be seen in Table 1B. [c]The most up to date version of CD-HIT is version 4.8 and therefore no version 5 beta could be found. [d]SEED could not run long sequences that were present in the Illumina data. [e]CD-HIT at 90% identity was not completed at time of publishing.

Table 1B: Uclust analysis of the five split files from SRR061270

Data file	**Time (minutes)**	**Clusters**
	Uclust '-id 0.95'	
xaa	142	1646610
xab	142	1647911
xac	142	1655233
xad	142	1657246
xae	16	637469
Total	**0**	**0**
	Uclust '-id 0.9'	
xaa	171	1557866
xab	171	1546910
xac	169	1546058
xad	171	1553121
xae	15	621747
Total	**0**	**0**

SRR061270.fasta was split at 10 million lines, or 5 million sequences, using the bash command "split -l 10000000 SRR061270.fasta".

Table 2A: Accuracy of Qiime on identifying OTUs

Data	**True OTU**	**Predicted OTU**	**Time (seconds)**
Divergent	23	18	62
Artificial	33	24	69
Even1	53	5	58
Even2	53	4	54
Even3	52	6	54
Uneven1	49	3	52
Uneven2	41	1	53
Uneven3	38	5	52
Titanium	69	59	71

Analyses were clustered by CD-HIT at 97% identity and then picked_open_reference_otus.py from Qiime was performed on the data.

Table 4A: Full clustering results of reference databases

Data	**Number of Sequences**	**Original file size**	**File size after clustering**	**Cutoff**	**Clusters**	**Reduced to (%)**	**Time (minutes)**
NCBI NR	18599335	9.9 GB	5.5 GB	90	10565347	44.4	533
16S (Silva + Greengene)	411839	641 MB	93 MB	98	57901	85.5	23
NCBI Bacteria							
Human body	1071335	325 MB	1.2 MB	90	3593	99.63	4
Human gut	817942	81 MB	115 KB	90	1086	99.86	1
Human body and gut	1889277	406 MB	1.3 MB	90	4679	99.68	7
NCBI Influenza	817587	1.4 GB	11 MB	95	6189	99.21	557

All analyses were done on 8 core 64 GB memory VM instances. Clustering was not performed on NCBI Bacteria.