Clustering Metagenome Sequences Using Canopies

Mohammad Arifur Rahman, Nathan LaPierre, Huzefa Rangwala and Daniel Barbara
Department of Computer Science
George Mason University
Fairfax, VA, United States
Email: mrahma23@gmu.edu, nlapier2@gmu.edu, rangwala@cs.gmu.edu and dbarbara@gmu.edu

***Abstract*—Advances in genome technologies has allowed for the collective sequencing of co-existing microbial communities (Metagenomics) that are ubiquitous across environments like the soil, ocean and human body. Metagenomics has spurred the development of several bioinformatics approaches for analyzing the diversity, abundance, function and role of the different organisms within these communities.**

We present a fast and scalable clustering algorithm for analyzing large-scale metagenome sequence data. Our approach achieves efficiency by partitioning the large number of sequence reads into groups (called canopies) using locality sensitive hashing. This initial approximate assignment of sequence reads to canopies is then refined by using state-of-the-art sequence clustering algorithms. This two-phase approach allows for the use of our developed algorithms as a pre-processing phase for computationally expensive clustering algorithms. Using the clusters as surrogates for Operational Taxonomic Units (OTUs) we estimate the biodiversity within a community sample. The Canopy-based clustering algorithm is evaluated on synthetic and real world 16S and whole metagenome benchmarks. We demonstrate the ability of our proposed approach to determine meaningful OTU assignments and observe significant speedup with regards to run time when combined with three different clustering algorithms.

***Keywords*-Clustering, Canopy, Metagenome, 16S, Biodiversity**

I. Introduction

A large portion of the earth's biomass comprises trillions of tiny microorganisms with varying biodiversity. Communities of interacting microbes exists in several ecosystems varying from the soil, ocean and human body [1]; and though, most of them are beneficial to the host some are known to cause unwanted conditions and can be linked to cause of diseases in the host [2]. *Metagenomics* is the sequencing of the collective DNA of microbial organisms coexisting as communities. Analyzing metagenomes has provided an unprecedented opportunity to understand the diversity, role and function of these organisms within different clinical and ecological environments [3][4].

However, sequencing technologies do not deliver the complete genome of an organism (millions in length), but large number of short contiguous subsequences called *reads* of length 75 to 500 in random order. Sequencing communities of genomes leads to additional challenges because reads from different microbes are mixed together; and upfront the composition of a community in terms of abundance and identity of microbial species is unknown [5]. Sequencing technologies also produce large datasets that range from Gigabytes (GB) to Terabytes (TB). Alternatively, targeted metagenome sequencing that involves sequencing of *marker genes* has been popular for the characterization of these communities. 16S and 18S sequences are repetitive marker genes within microbial genomes that are generally conserved but have slight variations in their sequences that allow them to be separated into different taxonomic groups [6].

In the past few years, several unsupervised clustering algorithms have been developed and used for the analysis of large-scale targeted and whole metagenome sequence reads (reviewed in Section II). The unsupervised grouping of similar sequences from the 16S/18S genes is referred as *binning* and allows for the estimation of biodiversity within a sample. Binning leads to assignment of similar sequences within groups called *Operational Taxonomic Units (OTUs)* [7]. Microbial OTUs are generally ecologically consistent across the hosts regardless of the clustering approaches [8].

In this paper we develop an efficient clustering algorithm based on Canopy Clustering [9] which can be used as a pre-clustering step with any state-of-the-art sequence clustering algorithm. Specifically, our approach identifies canopies with a greedy procedure and a fast sequence distance metric based on locality sensitive hashing [10]. Sequences within the canopies are considered as an initial partition (or grouping) of data which can be further refined by applying expensive and accurate clustering algorithms. Our developed approach treats each canopies independently allowing for easy parallelization.

We present experimental results on real and synthetic metagenome sequence benchmarks. We show that the use of Canopy Clustering (CC) in combination with UCLUST [11], SUMACLUST [13] and SWARM [12] leads to improved run time efficiency by 12.19, 21.09 and 18.61 times with accurate clustering results, respectively. We also demonstrate that our developed approach is scalable for large datasets with increased number of processors. The source code is freely available on Github[1] for public use.

[1]https://github.com/mrahma23/LSH-Canopy

II. Literature Review

In the past decade, several sequence clustering methods have been developed and used widely for metagenome sequences. A comprehensive survey by Kopylova et. al. [14] benchmarks various approaches including UCLUST [11], SWARM [12], SUMACLUST [13] and MOTHUR [15].

CD-HIT [16] is a general purpose sequence clustering algorithm that follows an incremental, greedy approach. CD-HIT uses pairwise sequence alignment to find similar sequences. UCLUST [11] is similar to CD-HIT but achieves a significant speedup over CD-HIT by using seeds (fixed length gapless subsequences) for performing pairwise sequence comparisons. MC-LSH [17] utilizes an efficient locality sensitive based hashing function to approximate the pairwise sequence similarity. MC-MinH [18] uses min-wise [19] hashing along with greedy clustering to group 16S and whole metagenome sequences. Mash [20], uses MinHash locality sensitive hashing to reduce large sequences to a representative sketch and estimate pairwise distances. Other methods for clustering sequence reads include TOSS [21], AbundanceBin [22] and CompostBin [23]. All unique kmers are first grouped in TOSS and then clusters are merged based on kmer repetitions. In AbundanceBin, reads are modeled as a mixture of Poisson distributions. Expectation Maximization (EM) algorithm is used to infer model parameters for the final clustering. Principal component analysis is used within CompostBin to project the data into a lower dimensional space followed with a graph partitioning approach. MOTHUR [15] uses a pairwise distance matrix as input and performs hierarchical clustering.

SWARM [12] uses exhaustive single-linkage clustering based on optimal sequence alignment. Sequences that are less than a certain distance from any other other sequence in the cluster are clustered together. SWARM attempts to reduce the impact of clustering parameters on the resulting OTUs by avoiding arbitrary thresholds and input sequence ordering dependencies. SWARM builds an initial set of OTUs by iteratively agglomerating similar amplicons. The amplicon abundance values are used to reveal OTUs internal structures and break them into sub-OTUs. SUMACLUST [13] is similar to UCLUST and uses a greedy strategy to incrementally construct clusters by comparing an abundance-ordered list of input sequences against the representative set of already-chosen sequences. UCLUST, SWARM and SUMACLUST are considered to be state-of-the-art metagenome sequence clustering methods by a benchmarking study [14]. As such, we compare the performance of our developed method with these three approaches.

III. Methods

A. *Overview*

Figure 1 shows an overview of our proposed canopy clustering algorithm. Sequence reads are represented with

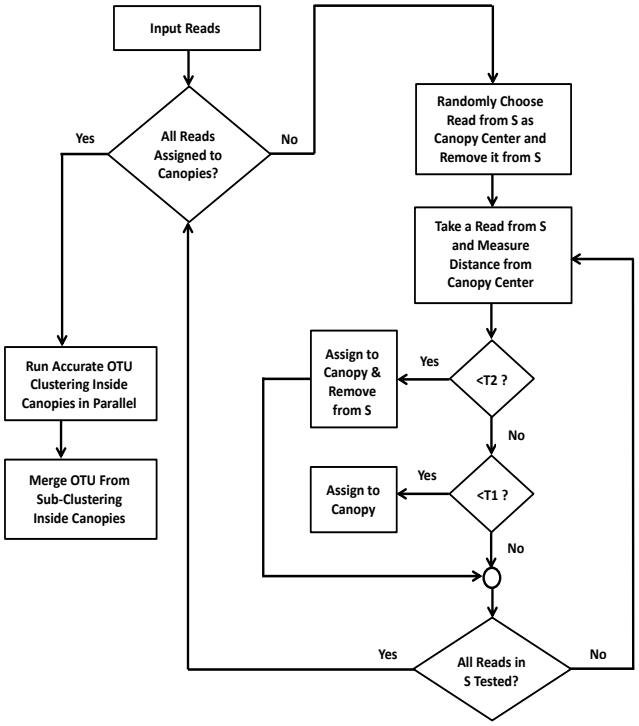


Figure 1. Workflow of Canopy Clustering for Large Scale Metagenome Data

kmers defined as contiguous subsequence of length k. One of the input reads is chosen randomly as the canopy centroid. Then we compute the distances of other reads to this canopy centroid. Two thresholds are used in Canopy clustering algorithm. If the distance is less than a *tight* threshold ($T2$) we assign the read to the canopy and that read will not be considered for another canopy. Otherwise, if a distance is less than a *soft* threshold ($T1$) we assign the read to the canopy but that read will be available for another canopy. This process continues until all sequence reads are assigned to atleast one canopy. For fast canopy assignment we use a random projection based technique. Specifically, we use Locality Sensitive Hash (LSH) to compute pairwise distances. More accurate clustering methods are used to sub-cluster each canopy in parallel. The sub-clustering method considers only the members within the canopy which significantly reduces pairwise distance computations.

B. *Canopy Clustering*

Canopy Clustering [9] is an efficient approximate clustering algorithm often used as pre-processing step for other accurate and expensive clustering methods like K-means or Hierarchical clustering. It is intended to speed up the clustering operations for large data sets, where standard clustering algorithms may be impractical due to run time and memory requirements. For a dataset with N instances, the

worst case calculations without canopy clustering is $O(N^2)$. Using canopy clustering, assuming the number of canopies is set to k; the worst case calculations with canopy clustering is $\sum_{i=1}^{k}(c_i)^2$ where c_i is the number of instances within i-th canopy.

Canopy clustering uses two distance thresholds, (i) *soft* threshold $T1$ and (ii) *tight* threshold $T2$. If data point p_1 is within the soft distance threshold $T1$ with centroid p_2 then p_1 will reside in same canopy as p_2 but p_1 may belong to other canopies assuming that it has only met soft threshold and best match is yet to be found. Thus one data point may belong to multiple canopies. However, if data point p_1 is within the tight distance threshold $T2$ with centroid p_2 then canopy clustering assigns p_1 to the same canopy as p_2 and stops assigning p_1 to any other canopy assuming that tight threshold has been met and best canopy assignment for p_1 has been found. Canopy centroids are selected randomly until all data points are assigned to at least one canopy.

C. *Locality Sensitive Hashing*

Canopies are intended to reduce pairwise distance calculations. As such, canopy clustering should be efficient which requires a fast and approximate distance measure. Locality Sensitive Hashing (LSH) [10] provides a solution for the approximate or exact near neighbors search problem. Given, a space S of points with a distance measure $d(x, y)$; a family H of hash functions is said to be $(d1, d2, p1, p2)$-sensitive for any x and y in S: if $d(x, y) < d1$, then the probability over all $h \in H$, that $h(x) = h(y)$ is at least $p1$ and if $d(x, y) > d2$, then the probability over all $h \in H$, that $h(x) = h(y)$ is at most $p2$. LSH provides a fast approximate distance measure while reducing data dimensionality making it appropriate for Canopy clustering for 16S and whole metagenomic data.

We construct the LSH family with bit sampling [24]. The normalized kmers frequency based feature vectors of sequence reads were first projected into d dimensional vectors in $\{0, 1\}^d$ space. Given an input vector v and a random hyperplane defined by r, we let $h(v) = \{0, 1\}$ based on $sgn(v \cdot r) = \pm 1$ that indicates on which side of the hyperplane v lies. Each possible choice of r defines a single function. Let H be the set of all such functions. For any $h_i \in H$ and for any two data x, y the probability that x and y *agree* on i^{th} positions of their respective d-length binary vector is

$$P[h_i(x) = h_i(y)] = 1 - \frac{distance(h_i(x), h_i(y))}{d} \qquad (1)$$

where *distance* is the hamming distance and d is the number of bits. Hence $H = \{h_1, h_2, ..., h_d\}$ is a $(d_1, d_2, 1 - d_1/d, 1 - d2/d)$ sensitive LSH family. The random projection and hamming distance calculation is computationally cheap and efficient making it suitable for fast partitioning of large volume of sequence reads.

D. *Sub-Clustering Inside Canopies*

Canopy clustering makes initial approximate partitions of the dataset and reduces pairwise distance computations. Each of these partitions can be further clustered in parallel and independendtly with expensive (but accurate) clustering methods. We use UCLUST [11], SUMACLUST [13] or SWARM [12] as accurate and expensive sub-clustering methods for the different canopies in this study.

E. *Merging results from Canopies*

Each cluster (OTU) is represented by the Longest Common Subsequence of all member sequences. The final step of our proposed framework is to merge the OTU representations generated by the canopies. According to Canopy cluster algorithm a single data point may belong to multiple canopies as long as the soft threshold is met. As a result similar OTU representations may appear from multiple canopies. To eliminate redundancy we run UCLUST on the OTU representations.

In this paper, the Canopy clustering (CC) approach integrated with UCLUST, SUMACLUST and SWARM are represented as CC_{UCLUST}, $\text{CC}_{SUMACLUST}$ and CC_{SWARM}, respectively.

IV. Experimental Evaluation

A. *Dataset Description*

To evaluate the performance of our proposed approach we use previously published synthetic and real world 16S, 18S and metagenome sequence benchmarks. Key statistics and relevant information regarding these datasets are presented in Table I.

Table I
Dataset Statistics

Datasets	**Type**	**# of Reads**	**# of Samples**	**Platform**
Bokulich_2 [26]	M	6,938,836	4	H
Bokulich_3 [26]	M	3,594,237	4	H
Bokulich_6 [26]	M	250,903	1	H
Canadian Soil [27]	R	2,966,053	13	H
Body Sites [28]	R	886,630	602	G
Global Soil [29]	R	9,252,764	57	H
Liver Cirrhosis [30]	R	6,117,828,130	232	H

Table shows information about dataset used in this study. M, R, H and G represent Mock, Real World, HiSeq and GS-FLX, respectively.

Table II
Parameter Settings

Datasets	Total # of Reads	# of Reads in Sampled Data	Parameters		
			d	$T1$	$T2$
Bokulich_2	6,938,836	693,884	47	0.48	0.36
Bokulich_3	3,594,237	359,428	31	0.43	0.34
Bokulich_6	250,903	25,090	17	0.37	0.21
Canadian Soil	2,966,053	296,605	29	0.46	0.37
Body Sites	886,630	88,663	22	0.39	0.22
Global Soil	9,252,764	925,276	59	0.49	0.38
Liver Cirrhosis	3,000,000	300,000	48	0.42	0.35
Liver Cirrhosis	30,000,000	3,000,000	67	0.51	0.39

Table III
PERFORMANCE COMPARISON [F-SCORE AND PEARSON CORRELATION COEFFICIENT (ρ)]

Methods	Comparison Metric	Datasets							
		Synthetic			*Real World*				
		Bokulich$_2$	Bokulich$_3$	Bokulich$_6$	Body Sites	Canadian Soil	Global Soil	Liver Cirrhosis Metagenome (Sampled 3M)	Liver Cirrhosis Metagenome (Sampled 30M)
UCLUST	F-Measure	0.39	0.40	0.51	N/A	N/A	N/A	N/A	N/A
CC$_{UCLUST}$	F-Measure	***0.39***	***0.41***	***0.52***	N/A	N/A	N/A	N/A	N/A
	(ρ) with Respect to UCLUST	***0.9831***	***0.9753***	***0.9831***	***0.9682***	***0.8419***	***0.9824***	***0.9216***	***0.9072***
SUMACLUST	F-Measure	0.40	0.41	0.51	N/A	N/A	N/A	N/A	N/A
CC$_{SUMACLUST}$	F-Measure	***0.41***	***0.42***	***0.51***	N/A	N/A	N/A	N/A	N/A
	(ρ) with Respect to SUMACLUST	***0.9709***	***0.9813***	***0.9538***	***0.9518***	***0.7643***	***0.8714***	***0.9281***	***0.8614***
SWARM	F-Measure	0.46	0.48	0.55	N/A	N/A	N/A	N/A	N/A
CC$_{SWARM}$	F-Measure	***0.46***	***0.49***	***0.56***	N/A	N/A	N/A	N/A	N/A
	(ρ) with Respect to SWARM	***0.9817***	***0.97861***	***0.9251***	***0.9648***	***0.7581***	***0.9143***	***0.9263***	***0.8533***

Table shows values of F-measures and Pearson Correlation Coefficient (ρ-value) of UCLUST, SUMACLUST, SWARM and their respective versions with Canopy clustering. F-measures is only available for synthetic datasets but not for real world datasets since no ground truths like known taxonomic profiles are available for them. ρ-value was calculated based on the taxonomy profiles at Genus level generated from clustered OTUs provided by a method and it's corresponding Canopy counterpart. Higher F-measures reflect better clustering by adhering to ground truth. Higher ρ-values reflect stronger correlation between taxonomic profiles. Higher F-scores and ρ values are represented with bold and italic letters.

Table IV
BIODIVERSITY COMPARISON [FAITHS PHYLOGENETIC DIVERSITY METRIC (PD), SHANNON AND SIMPSON]

Methods	Comparison Metric	Datasets						
		Synthetic			*Real World*			
		Bokulich$_2$	Bokulich$_3$	Bokulich$_6$	Body Sites	Canadian Soil	Global Soil	Liver Cirrhosis Metagenome (Sampled 3M)
UCLUST	PD Range	$[171.95 - 221.85]$	$[186.90 - 212.84]$	$[104.51 - 104.51]$	$[1.46 - 46.79]$	$[0.30 - 1352.73]$	$[2.98 - 3.29]$	$[3.14 - 51.47]$
	Shannon Range	$[2.52 - 3.51]$	$[2.43 - 3.54]$	$[5.87 - 5.87]$	$[0.29 - 7.67]$	$[2.32 - 7.86]$	$[1.84 - 8.30]$	$[0.27 - 5.77]$
	Simpson Range	$[0.55 - 0.75]$	$[0.55 - 0.76]$	$[0.96 - 0.96]$	$[0.049 - 0.98]$	$[0.80 - 0.99]$	$[0.0 - 0.98]$	$[0.02 - 0.97]$
CC$_{UCLUST}$	PD Range	$[164.79 - 217.72]$	$[169.41 - 198.36]$	$[109.33 - 109.33]$	$[2.37 - 47.13]$	$[0.52 - 1419.31]$	$[3.02 - 3.81]$	$[4.61 - 53.29]$
	Shannon Range	$[2.61 - 3.83]$	$[2.92 - 3.91]$	$[6.41 - 6.41]$	$[0.93 - 7.13]$	$[3.26 - 7.61]$	$[3.72 - 7.38]$	$[0.13 - 6.17]$
	Simpson Range	$[0.64 - 0.87]$	$[0.56 - 0.93]$	$[0.96 - 0.96]$	$[0.081 - 0.99]$	$[0.84 - 0.99]$	$[0.21 - 0.99]$	$[0.09 - 0.98]$
SUMACLUST	PD Range	$[106.00 - 162.78]$	$[142.85 - 174.19]$	$[89.22 - 89.22]$	$[0.93 - 39.47]$	$[0.59 - 1279.29]$	$[2.98 - 3.29]$	$[0.28 - 49.61]$
	Shannon Range	$[2.00 - 3.01]$	$[2.19 - 3.28]$	$[5.48 - 5.48]$	$[0.16 - 7.43]$	$[2.32 - 7.32]$	$[1.00 - 7.89]$	$[1.37 - 7.83]$
	Simpson Range	$[0.52 - 0.73]$	$[0.54 - 0.75]$	$[0.95 - 0.95]$	$[0.027 - 0.98]$	$[0.80 - 0.99]$	$[0.40 - 0.98]$	$[0.19 - 0.91]$
CC$_{SUMACLUST}$	PD Range	$[114.96 - 171.57]$	$[147.85 - 187.91]$	$[93.81 - 93.81]$	$[0.86 - 41.63]$	$[0.81 - 1292.34]$	$[1.37 - 4.89]$	$[0.18 - 51.35]$
	Shannon Range	$[2.51 - 3.94]$	$[2.96 - 4.11]$	$[5.94 - 5.94]$	$[1.21 - 7.13]$	$[3.12 - 7.79]$	$[2.17 - 7.25]$	$[1.91 - 7.39]$
	Simpson Range	$[0.68 - 0.79]$	$[0.51 - 0.74]$	$[0.96 - 0.96]$	$[0.06 - 0.99]$	$[0.88 - 0.99]$	$[0.23 - 0.98]$	$[0.27 - 0.88]$
SWARM	PD Range	$[18.37 - 24.73]$	$[17.36 - 19.81]$	$[30.84 - 30.84]$	$[1.44 - 28.66]$	$[0.54 - 706.57]$	$[5.79 - 6.18]$	$[2.06 - 39.71]$
	Shannon Range	$[2.98 - 3.91]$	$[2.01 - 3.04]$	$[5.03 - 5.03]$	$[0.28 - 7.63]$	$[1.0 - 7.79]$	$[1.66 - 7.81]$	$[1.47 - 6.91]$
	Simpson Range	$[0.70 - 0.82]$	$[0.53 - 0.74]$	$[0.95 - 0.95]$	$[0.05 - 0.98]$	$[0.50 - 0.99]$	$[0.00 - 0.99]$	$[0.18 - 0.89]$
CC$_{SWARM}$	PD Range	$[19.18 - 26.87]$	$[18.43 - 22.61]$	$[31.48 - 31.48]$	$[2.34 - 29.97]$	$[1.37 - 748.71]$	$[2.34 - 8.46]$	$[3.18 - 40.73]$
	Shannon Range	$[1.66 - 4.87]$	$[1.19 - 4.13]$	$[6.03 - 6.03]$	$[0.89 - 7.13]$	$[2.81 - 8.06]$	$[2.81 - 7.87]$	$[2.03 - 6.88]$
	Simpson Range	$[0.66 - 0.88]$	$[0.41 - 0.81]$	$[0.91 - 0.91]$	$[0.11 - 0.99]$	$[0.74 - 0.99]$	$[0.14 - 0.99]$	$[0.07 - 0.86]$

Table shows ranges of values for Faiths Phylogeny Diversity (PD), Shannon and Simpson coefficient over all samples in a dataset in the format $[minimum - maximum]$. Most of these datasets contain multiple samples and Alpha diversity metrics like PD, Shannon and Simpson values are generated for each of these samples separately. Biodiversity metric values changes significantly over samples e.g. diversity from hair samples and teeth cavity are supposedly different. So instead of mean values this Table represents $[minimum - maximum]$ ranges of values a sample can take. Similar ranges reflect similar diversity.

Synthetic Datasets:

1) Bokulich$_2$: This dataset was prepared using the Illumina TruSeq v2 paired–end library preparation kit. It is a simulated 16S rRNA gene microbial community dataset. This dataset contains 19 taxonomic Families, 19 Genera, 22 Species and 22 Strains in total. This dataset can also be found in the QIIME database (identifier 1685).

2) Bokulich$_3$: Similar to Bokulich$_2$ except that it was prepared with the TruSeq v1 paired-end library kit at Illumina Cambridge and is also available in the QIIME database (identifier 1686).

3) Bokulich$_6$: This 16S rRNA dataset was sequenced at Washington University School of Medicine and contains evenly distributed microbial communities. This dataset contains 13 taxonomic Families, 23 Genera, 44 Species and 48 Strains in total.

All these datasets from Bokulich et al. [26] are available at QIIME database[2] under their respective ID's. Since, these are simulated datasets the taxonomic profile of microbial organisms within them are known.

Real World Datasets:

[2]http://qiime.org/home_static/dataFiles.html

Table V
RUNTIME COMPARISON (IN MINUTES)

Datasets			Methods								
Type	Title	# of Reads	UCLUST	CC_{UCLUST}	Speed Up	SUMACLUST	$CC_{SUMACLUST}$	Speed Up	SWARM	CC_{SWARM}	Speed Up
Synthetic	$Bokulich_2$	6,938,836	12.71	***3.08***	***4.13x***	114.53	***13.89***	***8.24x***	128.12	***17.24***	***7.43x***
	$Bokulich_3$	3,594,237	10.43	***3.61***	***2.89x***	27.73	***5.11***	***5.43x***	18.37	***3.39***	***5.41x***
	$Bokulich_6$	250,903	4.47	***2.09***	***2.14x***	5.61	***2.07***	***2.71x***	4.17	***1.29***	***3.21x***
Real World	Body Sites	886,630	9.46	***3.03***	***3.12x***	18.42	***6.02***	***3.06x***	16.37	***5.83***	***2.81x***
	Canadian Soil	2,966,053	13.65	***4.17***	***3.27x***	363.96	***56.61***	***6.43x***	117.53	***20.84***	***5.64x***
	Global Soil	9,252,764	108.21	***18.75***	***5.77x***	510.92	***45.37***	***11.26x***	289.51	***34.84***	***8.31x***
	Liver Cirrhosis Metagenome	3,000,000	14.57	***4.03***	***3.61x***	46.62	***9.02***	***5.17x***	41.37	***8.75***	***4.73x***
	Liver Cirrhosis Metagenome	30,000,000	22.41h	***1.84h***	***12.19x***	46.62h	***2.21h***	***21.09x***	37.43h	***2.01h***	***18.61x***

4) Canadian Soil: The Canadian Soil dataset[3] contains genomic data of soil spanning from Arctic Tundra to Agricultural soil suitable for different agricultural products.

5) Body Sites: This dataset contains composition of bacterial communities from up to 27 different body sites in healthy adults. A collection of 602 samples acquired from different body sites of human subjects are provided with meta-data.

6) Global Soil: The global soil data was taken from Ramirez et al. [29] which is a study of the below-ground diversity in New York City's Central Park.

7) Liver Cirrhosis: The Liver Cirrhosis dataset was taken from the study by Qin et al. [30]. This is a whole gut microbiome wide association study of stool samples from 98 liver cirrhosis patients and 83 healthy controls to characterize the fecal microbial communities and their functional composition. In total, 860 GB of high-quality sequence data was generated in this study. Because of the high volume of sequence reads in this dataset, we used random sampling for feasible clustering performance evaluations. Two samples of 3 million and 30 million sequence reads were extracted from original dataset regardless of sample labels (disease or control) for our study.

B. *Evaluation Metrics*

We evaluate the performance of our developed clustering approach using the following commonly used metrics that are used for the assessment of (i) outputs from clustering algorithms, (ii) biodiversity within metagenome samples and (iii) computational run time.

1) Faiths phylogenetic diversity metric (PD): Faiths phylogenetic diversity [31] combines all the branch lengths of phylogenetic tree as a measure of diversity. So, if a new OTU is found and it is closely related to another OTU in the sample, it will contribute to a small increase to the PD score. However, if a new OTU from different lineage is found then it will contribute to a large increase in the PD score.

2) Shannon Entropy: Shannon-Wiener diversity index is defined as:

$$H = -\sum_{i=1}^{s} (p_i \log_2 p_i) \tag{2}$$

[3]http://www.cm2bl.org/

where s is the number of OTUs and p_i is the proportion of the community represented by OTU i. The Shannon index increases as both the richness and evenness of the community increase. The fact that the index incorporates both components of biodiversity can be seen as both a strength and a weakness. It is a strength because it provides a simple summary, but it is a weakness because it makes it difficult to compare communities that differ greatly in richness.

3) Simpson's Index: Simpsons index is defined as $1 - dominance$ or

$$1 - \sum p_i^2 \tag{3}$$

where where p_i is the proportion of the community represented by OTU i. Simpsons index is based on the probability that any two individuals drawn at random from an infinitely large community belong to the same species. It measures *evenness* of the community from 0 to 1. Higher value of this index implies higher similarity and relatively lower diversity of microorganisms within a sample.

4) F-measure: In case of synthetic datasets, expected taxonomic composition is known (*ground truth*). False-positive (FP) refers to the number of taxonomy that was found in observed but not expected, false-negative (FN) refers to the number of taxonomy that exists in expected but not observed, and true-positive (TP) refers to the number of taxonomy exists in both observed and expected. The following definitions were used:

$$precision = \frac{TP}{(TP + FP)} \tag{4}$$

$$recall = \frac{TP}{(TP + FN)} \tag{5}$$

$$FScore = \frac{2 \times precision \times recall}{(precision + recall)} \tag{6}$$

5) Pearson Coefficient Correlation (ρ-value): After getting Operational Taxonomic Units (OTU) from a clustering method we create a taxonomic profile at the Genus level. Pearsons correlation coefficient was computed to measure the relatedness of taxonomic assignment between a pair of tools. Values range between -1 and 1, with -1 indicating a negative correlation, 0 indicating no correlation, and 1 indicating a positive correlation or strong relationship.

C. Parameter Settings

For kmers, the value of parameter k was set to 4. The parameters for bit length of LSH (d), canopy's soft ($T1$) and hard ($T2$) thresholds were set by performing a grid search and validation. For validation purposes, 10% of the data was randomly sampled. First the parameter for LSH bit length d was estimated. In order to know the correctness of our estimation of d, we performed Canopy clustering with corresponding sub clustering methods inside theg canopies and compared the results corresponding to the bit lengths. For synthetic dataset, $F1$-scores and for real world datasets, the Pearson Coefficient Correlation (ρ) was compared with results from the corresponding expensive clustering approaches. We started with bit length d=1 and continued to increase value of d as long as we got better results for sampled data. Initial value of soft ($T1$) and tight ($T2$) threshold were set to 0.6 and 0.4, respectively. These are relaxed initial values for $T1$ and $T2$. as ithey allow for higher repetitions of sequence reads in multiple canopies due to higher value of $T1$ and easy final canopy assignment due to higher value of $T2$ which results in rough and fast approximate clusters prior to expensive sub-clustering step. Any bit length d that performs comparatively better results for these relaxed $T1$ and $T2$ values on sampled data is expected to bring better results for more strict $T1$ and $T2$ on whole datasets.

Once bit length of LSH projection (d) is selected we estimated values for $T1$ and $T2$ on the same sampled data. First, hard threshold ($T2$) was estimated with soft threshold ($T1$) fixed at 0.6. $T2$ determines which sequence reads will be retained for next iteration of canopy clustering. Same comparison metrics from the estimation of parameter d were used. We decreased the value of $T2$ from 0.4 (*relaxed*) to 0.1 (*strict*) with a step size of 0.01. The $T2$ corresponding to best result was selected. Then $T2$ was fixed at the best estimation and $T1$ was varied from 0.6 to best estimation for $T2$ with step size of 0.01. The value for $T1$ corresponding to best result was selected. Table II shows the estimated parameter values for our study.

D. Hardware and Software for Experiments

We performed all the experiments on computers with Intel 5th generation Core i7 2.70GHz 64bit processor with 8 core CPUs and 12GB memory. Sequence identity threshold for sub-clustering methods were set to default 97%. For implementation we used Python 2.7.12 and QIIME [32] version 1.9.0 (for diversity estimation). Taxonomy for reported OTUs was assigned using the RDP Classifier [33] against the 97% representative databases for Greengenes [34] and Silva [35] for methods used in this study. We used PyNast[4] open source sequence aligner for aligning clustered output.

[4]http://biocore.github.io/pynast/

V. Results

A. Runtime Comparison

Table V shows the run time in minutes for UCLUST, SUMACLUST, SWARM and their respective versions with our proposed Canopy clustering pipeline.

For the 30 million sampled data CC outperforms UCLUST, SUMACLUST and SWARM by 12.19x, 21.09x and 18.61x respectively. For the other larger dataset titled Global Soil CC outperforms UCLUST, SUMACLUST and SWARM by 5.77x, 11.26x and 8.31x respectively. The highest gains in runtime in our study were found for the largest dataset which is the 30 million sampled data. This indicates that CC scales well with large scale data.

B. Effect of Varying Number of Processors

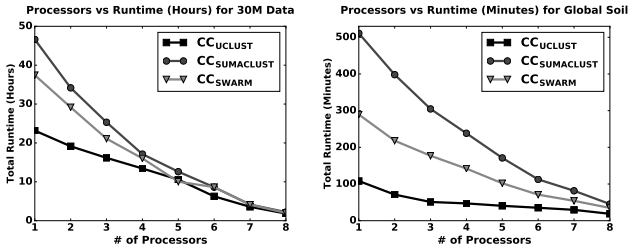
Figure 2a-2b shows the runtime of CC_{UCLUST}, $\text{CC}_{SUMACLUST}$ and CC_{SWARM} on two largest benchmarks used in this study - Global Soil and the 30M Liver Cirrhosis dataset. We observe that increasing the number of processors reduces the total runtime. Significant reductions in run time were observed $\text{CC}_{SUMACLUST}$ and $\text{CC}_{SUMACLUST}$ as compared to CC_{UCLUST}.

C. Clustering Performance

Table III shows the performance of UCLUST, SUMACLUST and SWARM in comparison to CC clustering versions. Table III compares F-scores and Pearson Correlation Coefficient (ρ). F-scores are only available for synthetic benchmarks since taxonomic profile for them is known. Correlation values were generated based on taxonomic profiles at the Genus level, generated from outputs of the clustering methods. We observe that F-scores obtained from a clustering method and its corresponding Canopy clustering version are very similar and in some cases better. We see a higher F-score for Bokulich_2 benchmark from $\text{CC}_{SUMACLUST}$ in comparison to SUMACLUST. CC_{UCLUST} and CC_{SWARM} resulted in the same F-scores as their respective naive versions for Bokulich_2. For Bokulich_3 benchmark we observed higher F-scores for all Canopy clustering methods. Finally for Bokulich_6 benchmark, the F-scores of CC_{UCLUST} and CC_{SWARM} were improved comparing to UCLUST and SWARM, respectively. For all benchmarks we observe strong correlations between taxonomic profiles at genus level. The highest correlation was observed for Bokulich_6 benchmark between UCLUST and CC_{UCLUST} with 0.9831.

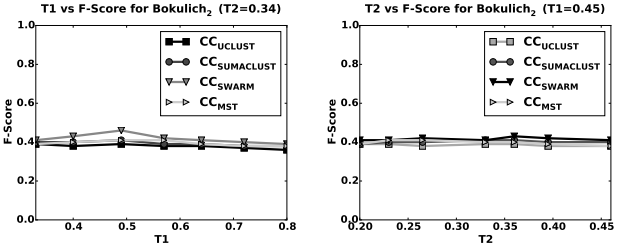
D. Biodiversity Estimation

Clustering metagenome sequences outputs OTUs that represent biodiversity contained in the samples. Table IV shows Faiths phylogenetic diversity metric (PD), Shannon and Simpson index after clustering with different methods. These metrics are popular Alpha Diversity metrics that measure species diversity in sites or habitats at a local scale.



(a) (b)

Figure 2. Effect of number of processors on Runtime of CC_{UCLUST}, $CC_{SUMACLUST}$ and CC_{SWARM} for 30M sampled dataset and Global Soil dataset, two of the largest datasets used in this study.



(a) (b)

Figure 3. 3a and 3b show effect of varying $T1$ and $T2$ on F-scores for the largest synthetic dataset Bokulich$_2$

These metric values are generated per sample basis. Table IV shows ranges of metric values in $[minimum - maximum]$ format for the different methods with and without Canopy clustering. Any sample of a dataset will take value from this range. These ranges may not be same since OTUs vary over clustering methods. From Table IV, we observe that Canopy clustering based methods produce similar ranges of values as their naive counterparts. No significant changes in diversity metric values were observed which indicates that the Canopy based approach reproduces similar biodiversity estimates while reducing run time.

E. Sensitivity Analysis (Varying T1 and T2)

Figure 3a-3b shows the effect of varying $T1$ ($T2$ fixed at 0.34) and $T2$ ($T1$ fixed at 0.45) on F-scores for the Bokulich$_2$ dataset. Reducing $T1$ leads to comparatively *strict* soft-threshold which will reduce repetitions of instances in multiple canopies. For a fixed $T2$ this implies that lower $T1$ will yield better canopy assignment. From Figure 3a we can say that when $T1$'s range is in 0.4 to 0.6 our proposed approach provides better F-Scores. On the other hand, increasing $T2$ leads to comparatively *relaxed* tight-threshold for canopies. As a result instances will be prematurely assigned to canopies without waiting for best match. From Figure 3b we note that when $T2$' is in between 0.25 to 0.35, our proposed approach achieves better F-Scores. Lowering $T2$ may lead to higher run time since instances will continue to reappear until the $T2$ threshold is satisfied.

VI. Conclusion and Future Work

We developed a greedy approximate clustering process for any accurate and relatively expensive clustering on large scale metagenome datasets. Our approach takes advantage of the multi-core CPU systems by partitioning the large dataset with a fast and cheap pairwise distance measure and then deploying comparatively expensive clustering in parallel which considers only data points that are inside a partition. Our proposed approach scales well with large datasets and provide significant reduction in computation time. We demonstrate that our approach provides similar outcome in terms of biodiversity metrics, ground truth and taxonomic correlation with corresponding expensive clustering methods.

References

[1] Turnbaugh et al., "The human microbiome project: exploring the microbial part of ourselves in a changing world," *Nature*, vol. 449, no. 7164, p. 804, 2007.

[2] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project: exploring the microbial part of ourselves in a changing world," *Nature*, vol. 449, no. 7164, p. 804, 2007.

[3] Qin et al., "A human gut microbial gene catalogue established by metagenomic sequencing," *nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[4] M. Pop and S. L. Salzberg, "Bioinformatics challenges of new sequencing technology," *Trends in Genetics*, vol. 24, no. 3, pp. 142–149, 2008.

[5] H. Teeling and F. O. Glöckner, "Current opportunities and challenges in microbial metagenome analysisa bioinformatic perspective," *Briefings in bioinformatics*, p. bbs039, 2012.

[6] Chakravorty et al., "A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria," *Journal of microbiological methods*, vol. 69, no. 2, pp. 330–339, 2007.

[7] Blaxter et al., "Defining operational taxonomic units using dna barcode data," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1462, pp. 1935–1943, 2005.

[8] T. S. Schmidt, J. F. M. Rodrigues, and C. Von Mering, "Ecological consistency of ssu rrna-based operational taxonomic units at a global scale," *PLoS Comput Biol*, vol. 10, no. 4, p. e1003594, 2014.

[9] McCallum, Andrew and Nigam, Kamal and Ungar, Lyle H., "Efficient clustering of high-dimensional data sets with application to reference matching," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '00. New York, NY, USA: ACM, 2000, pp. 169–178.

[10] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proceedings of the 25th International Conference on Very Large Data Bases*, ser. VLDB '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 518–529.

[11] Edgar, Robert C., "Search and clustering orders of magnitude faster than blast," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.

[12] Mahé et al., "Swarm v2: highly-scalable and high-resolution amplicon clustering," *PeerJ*, vol. 3, p. e1420, 2015.

[13] C, Mercier, F, Boyer, A, Bonin, and E, Coissac, "Sumatra and sumaclust: fast and exact comparison and clustering of sequences [submitted for publication]," *Available at https://git.metabarcoding.org/obitools/sumatra/wikis/home*, 2014.

[14] Kopylova et al., "Open-source sequence clustering methods improve the state of the art," *mSystems*, vol. 1, no. 1, 2016.

[15] Schloss et al., "Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and Environmental Microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[16] Li, Weizhong and Godzik, Adam, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[17] Z. Rasheed, H. Rangwala, and D. Barbara, "Efficient clustering of metagenomic sequences using locality sensitive hashing." in *SDM*. SIAM, 2012, pp. 1023–1034.

[18] Z. Rasheed and H. Rangwala, *MC-MinH: Metagenome Clustering using Minwise based Hashing*, ch. 74, pp. 677–685.

[19] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher, "Min-wise independent permutations," *Journal of Computer and System Sciences*, vol. 60, no. 3, pp. 630–659, 2000.

[20] Ondov et al., "Mash: fast genome and metagenome distance estimation using minhash," *bioRxiv*, p. 029827, 2016.

[21] O. Tanaseichuk, J. Borneman, and T. Jiang, "Separating metagenomic short reads into genomes via clustering," *Algorithms for Molecular Biology*, vol. 7, no. 1, p. 1, 2012.

[22] Wu, Yu-Wei and Ye, Yuzhen, "A novel abundance-based algorithm for binning metagenomic sequences using l-tuples," *Journal of Computational Biology*, vol. 18, no. 3, pp. 523–534, 2011.

[23] S. Chatterji, I. Yamazaki, Z. Bai, and J. A. Eisen, "Compostbin: A dna composition-based algorithm for binning environmental shotgun reads," in *Annual International Conference on Research in Computational Molecular Biology*. Springer, 2008, pp. 17–28.

[24] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," pp. 604–613, 1998.

[25] A. Z. Broder, "On the resemblance and containment of documents," pp. 21–29, 1997.

[26] Bokulich et al., "Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing," *Nat Methods*, vol. 10, pp. 57–59, 2013.

[27] Neufeld et al., "Open resource metagenomics: a model for sharing metagenomic libraries," *Standards in Genomic Sciences*, vol. 5, pp. 203–210, 2011.

[28] Costello et al., "Bacterial community variation in human body habitats across space and time," *Science*, vol. 326, no. 5960, pp. 1694–1697, 2009.

[29] Ramirez et al., "Biogeographic patterns in below-ground diversity in new york city's central park are similar to those observed globally," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 281, no. 1795, 2014.

[30] N. et al., "Alterations of the human gut microbiome in liver cirrhosis," *Nature*, vol. 513, no. 7516, pp. 59–64, 2014.

[31] Faith, Daniel P, "Conservation evaluation and phylogenetic diversity," *Biological conservation*, vol. 61, no. 1, pp. 1–10, 1992.

[32] Caporaso et al., "Qiime allows analysis of high-throughput community sequencing data," *Nat Meth*, vol. 7, no. 5, pp. 335–336, 2010.

[33] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy," *Applied and environmental microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.

[34] DeSantis et al., "Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb," *Applied and environmental microbiology*, vol. 72, no. 7, pp. 5069–5072, 2006.

[35] Pruesse et al., "Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb," *Nucleic acids research*, vol. 35, no. 21, pp. 7188–7196, 2007.