MetaCluster: Unsupervised Binning of Environmental Genomic Fragments and Taxonomic Annotation

Bin Yang[1, 2], Yu Peng[2], Henry C.M. Leung[2], S.M. Yiu[2],
Junjie Qin[3], Ruiqiang Li[3], Francis Y.L. Chin[2]
[1]State Key Laboratory of Bioelectronics, Southeast University, Nanjing, China
[2]Department of Computer Science, The University of Hong Kong, Hong Kong SAR, China
[3]BGI-Shenzhen, Shenzhen, China
{byang, ypeng, cmleung2, smyiu, chin}@cs.hku.hk; {qinjj, lirq}@genomics.org.cn

ABSTRACT
Limited by the laboratory technique, traditional microorganism research usually focuses on one single individual species. This significantly limits the deep analysis of intricate biological processes among complex microorganism communities. With the rapid development of genome sequencing techniques, the traditional research methods of microorganisms based on the isolation and cultivation are gradually replaced by metagenomics, also known as environmental genomics. The first step, which is also the major bottleneck of metagenomic data analysis id the identification and taxonomic characterization of the DNA fragments (reads) resulting from sequencing a sample of mixed species. This step is usually referred as "binning".

Existing binning methods based on sequence similarity and sequence composition markers rely heavily on the reference genomes of known microorganisms and phylogenetic markers. Due to the limited availability of reference genomes and the bias and unstable of markers, these methods may not be applicable in all cases. Not much unsupervised binning methods are reported, but the unsupervised nature of these methods makes them extremely difficult to annotate the clusters with taxonomic labels. In this paper, we present MetaCluster 2.0, an unsupervised binning method which could bin metagenomic sequencing datasets with high accuracy, and also identify unknown genomes and annotate them with proper taxonomic labels. The running time of MetaCluster 2.0 is at least 30 times faster than existing binning algorithms.

MetaCluster 2.0, and all the test datasets mentioned in this paper are available at http://i.cs.hku.hk/~alse/MetaCluster/.

Categories and Subject Descriptors
J.3. [**Computer Applications**]: Life and Medical Sciences – *biology and genetics.*

General Terms
Algorithms, Experimentation, Measurement, Performance,

Reliability, Verification.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Conference'10, Month 1–2, 2010, City, State, Country.
Copyright 2010 ACM 1-58113-000-0/00/0010…$10.00.

Keywords
Metagenomics, Binning, DNA composition features, *l*-mer,

Spearman Distance, taxonomic annotation, *k*-mean clustering.

1. INTRODUCTION
Traditional genomic study usually focuses on one single individual species (e.g. human). However, researchers have discovered that all the microorganisms present in a specific habitat have critical effects on one another and the host. For example, the unbalance or abnormal diversity of microbes in human is proved to be associated with common diseases such as Inflammatory Bowel Disease (IBD) [1] and gastrointestinal disturbance [2]. In particular, understanding the effects of microbial community on human may contribute to better diagnosis, prevention, and treatment of diseases. Genomic analysis on the collective genomes of all microorganisms from an environmental sample (also known as metagenomics, environmental genomics, or community genomics) becomes necessary. The difficulty of metagenomics lies on the fact that most of the species (can be up to 99%) found in a sample are unknown and cannot be easily cultivated and separated in a laboratory [3]. One possible approach is to make use of high-throughput sequencing technology to obtain DNA fragments (contigs) of different genomes from the mixed sample and perform analysis on the fragments [4]. Examples of metagenomics projects include Acid Mine Drainage Biofilm (AMD) which analyzes dozens of species [5] and the recent Human Gut Microbiome (HGM) which involves thousands of species [6].

Fragments of a metagenomics project are from multiple genomes for which most of them are unknown. The first step to analyze these fragments is to assign them to the taxonomy tree (referred as *binning*) [7] to obtain a general map and approximate taxonomic annotation of the microbe distribution of the sample. Depending on the research requirements, the quality and the complexity of the metagenomic sequencing (MS) dataset, the binning process could be done at various taxonomic levels from *Kingdom* (the highest) to some low levels such as *Genus*.

Existing binning methods could be roughly classified into three categories: sequence similarity-based method, sequence composition-based method and unsupervised-based method. Sequence similarity-based methods [8] try to align each DNA fragment to known reference genomes. Based on the alignment results (e.g. BLAST hits or selected phylogenetic specific marker genes [9]), each fragment is assigned to the taxonomic class with a reference genome showing a high similarity to the fragment. Since less than 1% of microorganisms can be cultured and

sequenced, many genomes of microorganisms are not in the database and many DNA fragments cannot be aligned well. Besides, a single run of sequencing machine can produce billion of DNA fragments [1] which require several days to align them with all known reference genomes.

Since the sequence similarity-based methods are time-consuming and only work well for the DNA fragments from species with known genomes, other researchers introduced sequence composition-based methods which cluster DNA fragments in a supervised or semi-supervised manner using generic features such as genome structure or composition. Structure features such as composition features of reference genomes or taxonomic marker regions (e.g. *16S rRNA*[10], *recA* and *rpoB* are commonly accepted fingerprint genes) are extracted. These generic features can be used to construct a classifier [11] for determining DNA fragments from different species or can be used as constraints for semi-supervised clustering or classification. These methods suffered from low availability and reliability of taxonomic markers. For example, study on several metagenomic projects, such as the enhanced biological phosphorus removing (EBPR) sludge [12], Sargasso Sea [4] and the Minnesota soil samples [13], indicated that only 0.17%, 0.06% and 0.017% of the contigs (DNA fragments) respectively are known to carry *16S rRNA* markers. Even if we select more markers such sas *recA* and *rpoB*, still less than 1% of the fragments could be identified. The reliability of taxonomic markers has also been challenged, some papers [14] reported that some species may share multiple markers with other species or multiple kinds of *16S rRNA* molecules exist in a single bacterium due to high mutation and gene exchange ratio of microbe, which leads to incorrect classifications.

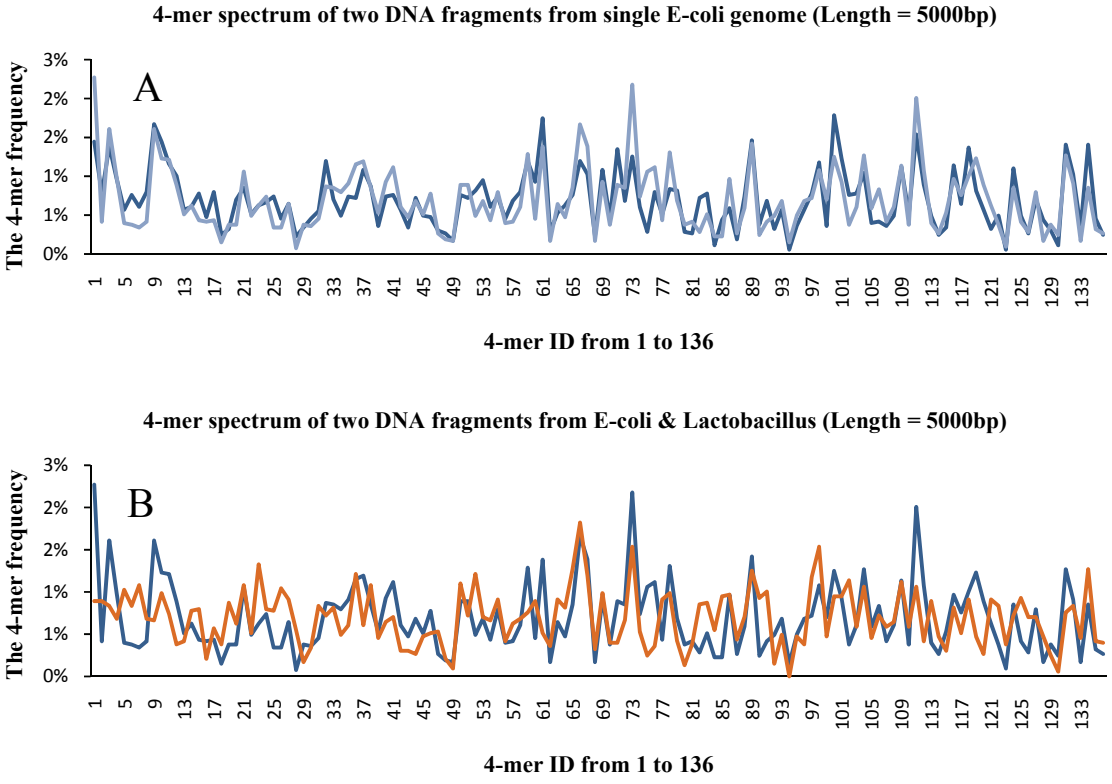

Figure 1. The 4-mer frequency spectrum. Figure. 1A is the 4-mer spectrums of two DNA fragments from the same E-coli genome. Figure. 1B is the 4-mer spectrums of two DNA fragments from the genomes of E-coli and Lactobacillus which belong to the same kingdom but different phyla.

Since there are not enough reference genomes and the generic features are not reliable, another direction is to consider unsupervised method for clustering DNA fragments based on the occurrence frequencies of the *l*-mer (short DNA substrings of length *l*) distribution of the DNA fragments [15,16]. In these approaches, each fragment can be regarded as a vector containing the occurrence frequencies of all possible *l*-mers in the fragment. The rationale behind is based on the observation that the *l*-mer distributions within same genome are more similar than the *l*-mer distributions of two unrelated species, say in different phyla. Figure 1 gives a straightforward about this phenomenon. This observation is supported by [17] which reported that the distribution of dinucleotide is more or less the same for the same genome, but varies between genomes from different taxonomic groups. TETRA [18,19] was the first who applied *l*-mer frequency distribution to the binning problem of metagenomic datasets. Their method requires long DNA fragments (40kb) in order to produce reasonable results. However, assembling reads from metagenomics data to long contigs is still not feasible, thus their method is not very practical. [20] Further investigates this approach and has come up with an approach to bin the fragments based on a carefully selected subset of *l*-mers. Although the

results in [20] are good, it is not clear how to select the subset of *l*-mers. Both the above papers do not provide a taxonomy annotation for the fragments. MEGAN [8] is the only tool that can provide taxonomy annotation for DNA fragments, however the core of MEGAN is based on the alignment of the fragments on known reference genomes. For fragments from unknown species, the results are not satisfactory. Recently, there is another unsupervised clustering method called LikelyBin [22] which makes use of Markov Chain Monte Carlo approach to model the genome sequences of different species. The model is complicated. Thus, it may have the problem of over fitting and requires a lot of computation to perform the clustering.

In this paper, we provide a two-step approach to solve the binning problem of the metagenomic data. We first cluster the fragments using an unsupervised approach (i.e., we do not make use of any known reference genomes) based also on the *l*-mer distribution of the fragments. Instead of using a selected subset of *l*-mers as in [20], we use all *l*-mers and apply the well known statistical measure, Spearman Footrule Distance [21], to capture the similarity of *l*-mer distributions of any two fragments. Combining with *k*-mean clustering algorithm, we group the fragments into the same cluster which a strong similarity on their *l*-mer distributions. Spearman Footrule Distance considers the relative ranking of the occurrence frequency of an *l*-mer compared to other *l*-mers which provides a more reasonable assessment on the similarity than using the absolute or normalized occurrence frequencies as in [20]. Our approach is simple and much faster (30-50 times faster than the best existing approach LikelyBin) and the average accuracy is at least as good as theirs.

More important, unlike all other unsupervised approaches which do not provide any taxonomic annotation to the clusters, our second step is to label (annotate) the clusters with taxonomic information even if some of the genomes are unknown. We believe that this step is important. Even providing an approximated annotation at high taxonomic ranks such as Family or Order helps the biologists to design follow-up experiments for further investigation. In order to assign the clusters to taxonomy tree, we represent each reference genome by a composition feature vector. The process of assigning the cluster to the taxonomic tree is similar to the most prevalent single-winner plurality (also called "first-past-the-post") voting system. Each DNA fragment in a cluster will vote for the nearest reference genome based on the Spearman Footrule Distance. Then according to the requested taxonomic rank, say genus (or higher level like family and order), the total number of votes for the reference genomes of the same genus will represent the support of this genus. Finally, the cluster will be assigned to the majority genus such is the winner of this voting. Experimental results demonstrated that we can assign the clusters to the taxonomy tree with high accuracy (about 87.5% to 91.85%) at different taxonomic ranks and is about 20% higher than MEGAN. We also show that our approach is robust in view of the amount of sequencing errors and relative species abundance ratios in the sample.

2. METHOD

As shown in Figure 2, our binning approach could be divided into two major steps. During the first step, an unsupervised *K*-mean clustering method assigns the mixed DNA fragments from several unknown species into clusters based on the similarity of their *l*-mer distributions. With the assumption that each cluster contains the DNA fragments from the same genome, these clusters are classified into the most similar taxonomic groups in the second step. So generally speaking, our solution identifies the unknown genomes from the metagenomic sequencing datasets (step 1) and provides taxonomic labels for the unknown genomes (step 2).

2.1 *l*-mer frequency calculation

The *DNA composition features* of each DNA fragment are represented by its *l*-mer frequencies. As there are 4 different DNA nucleotides, there are at most 4^l *l*-mers in a DNA sequence. If a sliding window of length *l* is slid along each DNA fragment and the frequency of every *l*-mer, say f_i, $i \in [1, 4^l]$, were recorded, then the total number of *l*-mers in a DNA fragment would be $\sum_{j=1}^{4^l} f_j$. For example, a DNA fragment of length 500bp has 497 4-mers and a DNA fragment of length 2000bp has 1997 4-mers. Thus the DNA is a *feature vector* defined as $[f_1, f_2 \cdots f_{4^l-1}, f_{4^l}]$. As each DNA fragment can be obtained from either strand of the DNA genome, the frequency of one *l*-mer and its reverse complement *l*-mer can be combined into a single frequency and this process will reduce the size of vector by half, i.e. $N(l) = 4^l/2$, if *l* is odd; $\left(4^l + 4^{l/2}\right)/2$, if *l* is even.

For example, in order to be effective and to have a reasonable vector size, *l* can be set to 4. So each DNA fragment will be represented by a feature vector with 136 components and the input metagenomic sequencing dataset of FASTA format will be transformed to a $n \times 136$ matrix with *n* rows representing *n* DNA fragments.

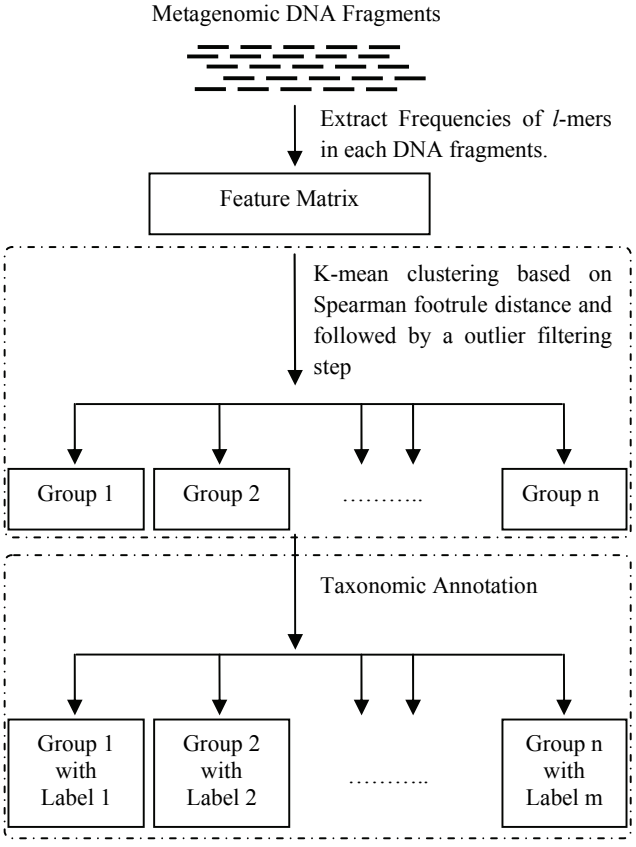

Figure 2. The pipeline of our method could be generally divided into two major steps: clustering and annotation.

2.2 Distance definition based on the ranked list correlation coefficient

Our binning method is based on a widely accepted observation [17] that the *l*-mer distributions of those DNA substrings (fragments) from the same genome are similar (see Figure 1). As the feature vector represents the *l*-mer distribution of a given DNA fragment, the similarity between two DNA fragments can be measured by the "distance" between their *l*-mer feature vectors. In [20], an "essential *l*-mer region" is selected to calculate the distance between two DNA fragments. This distance definition filters out both intra-species and inter-species noise and achieves reasonably good performance. However, this distance definition is very sensitive to the variation of *l*-mer occurrence frequencies and affects the stability of the performance. To improve the binning performance, we apply a commonly used correlation coefficient based distance definition called Spearman Footrule Distance which considers the relative ranking of the occurrence frequency of an *l*-mer with other *l*-mers.

Consider two DNA fragments A and B with the following 4-mer feature vectors:

A: $(a_1, a_2, \cdots, a_i, \cdots a_j, \cdots a_k)$

B: $(b_1, b_2, \cdots, b_i, \cdots b_j, \cdots b_k)$

The Spearman Footrule Distance is a very intuitive definition for comparing two ordered lists. Let $r^A(a_i)$ be the rank of a_i in the sorted list and $r^B(b_i)$ be the rank of b_i in the sorted list. Then the Spearman Footrule Distance is defined as:

$$Distance_s(A,B) = \sum|r^A(a_i) - r^B(b_i)|$$

The smaller the value of the metric, the more similar the vectors are. For vectors with size *k*, when the two vectors have no element in common, the maximum distance value is $k(k+1)$.

There is another widely used distance definition, Kendall's Tau Distance [23], whose clustering performance is very similar to the Spearman Footrule. However, since the computational complexity of computing Kendall's Tau Distance is higher, the Spearman Footrule Distance is used.

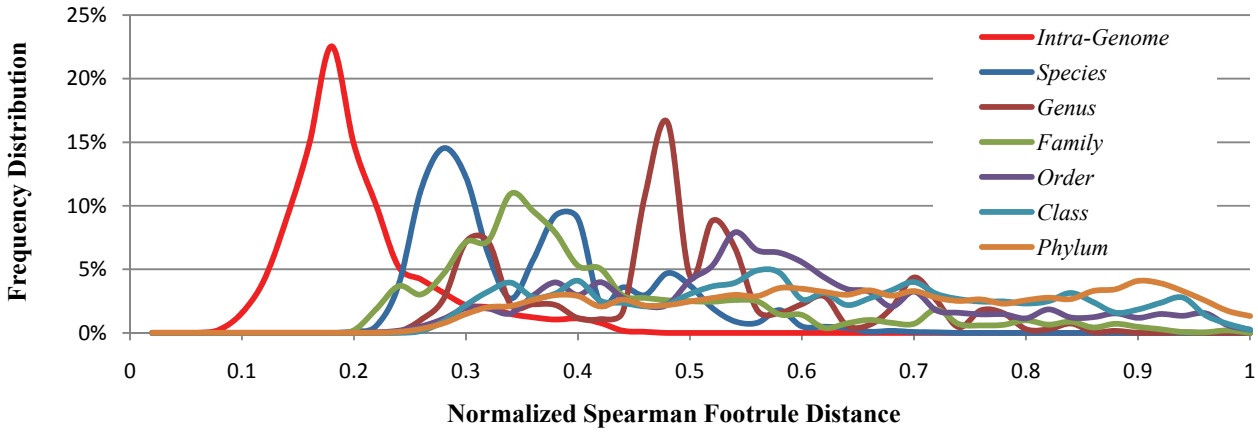

Figure 3. The frequency distribution of the intra-genome Spearman Footrule Distance and inter-genome Spearman Footrule Distance. The inter-genome distances are separated into 6 individual curves which represent 6 different taxonomic levels

2.3 Clustering and optimization

Based on the Spearman Footrule Distance, we confirm experimentally that the *l*-mer frequencies feature vectors of DNA fragments from the same genome tend to have similar distributions.

We randomly select 10,000 pairs of DNA fragments from each reference genome from the NCBI genome database, calculate the Spearman Footrule Distance for each pair, and then apply the average value of these 10,000 distances as the intra-genome distance for each genome. For all the genomes (1140 genomes in total) in the NCBI reference database, we plot the statistical distribution of these intra-genome distances in Figure 3.

We also test the inter-genome Spearman Footrule Distance between two randomly selected genome A and B at some specific different taxonomic differentia levels say, Species (which means genome A and B are from the same Genus but different Species), Genus (which means genome A and genome B are from the same Family but different Genus) as well as level like Family, Order, Class and Phylum. For each level, we select 1000 genome pairs, and for each genome pair, we select 10,000 pairs of DNA fragments, one from genome A and the other from B. Then we calculate the Spearman Footrule Distance of these 10,000 DNA fragments pairs and apply the average value as the inter-genome distance between genome A and B. So for each taxonomic differentia level we have 1000 inter-genome distances. Similar to the intra-genome distance, we plot the statistical distribution of these inter-genome distances in Figure 3. The intra-genome distance is consistently and significantly smaller than any other inter-genome distances no matter at which taxonomic differentia level.

Another observation is that, the *l*-mer feature vectors from the same genome tend to be located around the same cluster center. Based on this observation, we can simply apply the simple *K*-mean algorithm to cluster the fragments.

Suppose that we want to cluster the *l*-mers feature vectors of fragments into *K* clusters. Based on our distance definition, the objective function of *K*-mean is:

$$MinE = \sum_{i=1}^{K} \sum_{x \in C_i} Distance_s{}^2(x, c_i)$$

The vector c_i represents the center of cluster C_i.

The *K*-mean clustering algorithm is described in the Appendix (see Algorithm 1).

Because of the unstable feature of *K*-mean caused by the random selection of the initial clustering centers, we will run the algorithm with different initial clustering centers and choose the clustering result with the minimum objective value.

In [20], it is mentioned that even for the same genome, the *l*-mer distribution of some special genome region (such as promoters and exogenous transferred regions) can be very unique when compared with general genome regions. These *l*-mer data points could be considered outliers and they might introduce negative effects during the clustering process. So after the *K*-mean clustering, an additional outlier filtering step is introduced to improve the performance. So during the *K*-mean clustering process, we calculate the distance between every data point to the cluster, their average center distance μ and the standard deviation σ for each cluster. Those data points with center distance larger than $\mu + 2\sigma$ should be removed as outliers (Algorithm 2 in the Appendix).

Normally, after this step usually no more than 5% of the data points will be removed as outliers. Based on the experiment results, the average clustering accuracy will be increased by about 2% when compared with the performance before filtering.

2.4 Taxonomic annotation of clusters

After the first step of clustering, we have high confidence that the sequences in a cluster should come from the same genome. The exciting part of the unsupervised binning method is its potential to identify new genomes which have never be discovered before. On the other hand, unless we could provide a general description of the context in the clusters, the binning results remain helpless for following up practical researches. Without any background taxonomic information, it is extremely difficult for an unsupervised method to annotate the clusters with taxonomic labels. This deficiency seriously limits the practical application of the unsupervised binning methods. Although this objective is important, all the existing unsupervised binning methods seldom mention the requirement of annotation for the clusters generated after binning process.

If the genome exists in the NCBI taxonomic tree, it should be easy to classify the clusters by an alignment-based binning method. Since the input DNA fragments for our research are from unknown species which have no reference genomes in the NCBI database, the alignment-based method cannot be used for classifying the clusters. However, for the unknown species, there is high probability that the genome sequence of some similar species in the level of genius, family or higher ranks may be available in the NCBI taxonomic tree. We can annotate unknown genomes by estimating the similarity between the clusters and the reference genomes. Note that, even with this additional step of annotation, our approach quite different from the alignment or machine learning based binning methods. The main difference is that our approach directly identifies the unknown genomes and then annotates them with taxonomic information, instead of directly assigning the DNA fragments to the high level taxonomic groups. When compared with our approach whose resolution is performed at the genome level, the resolution of other approaches at a higher level is more ambiguous.

After the first clustering step, our second step of annotation can be described as follows:

Input:

Clusters of DNA fragments

Partial NCBI taxonomic tree (the nodes with complete genomes)

The taxonomic level determined by user to annotate the clusters

Output:

The clusters with particular taxonomic annotations

Here we introduce a voting method to annotate the clusters of unknown genomes with taxonomic labels. The DNA fragments from a cluster can be taken as “voters” and the reference genomes in the NCBI database as “candidates” belonging to several “political parties”. But instead of directly selecting the “candidates”, our objective is to select the “political party” which most voters agree with. In our case, the “party” is the majority taxonomic group to annotate the cluster. So the most prevalent single-winner plurality (also called "first-past-the-post") voting system is adopted in our approach. Each DNA fragment in a cluster will vote for the nearest reference genome based on the Spearman Footrule Distance. Then according to the requested taxonomic rank, for example genus (or higher level like family and order), the votes of the reference genomes from the same genus will be summed up to represent the votes of this genus. Finally, the cluster will be annotated with the winner of this election, i.e. the same taxonomic label as the majority genus. (The algorithm of this voting strategy is shown as Algorithm 3 in the Appendix).

3. RESULTS AND CONCLUSION

In this section, we analyze the performance of our binning algorithm, MetaCluster 2.0, from two aspects, the accuracy of clustering DNA fragments from same kind of species and the accuracy of annotating DNA fragments to the taxonomic tree under different situations. We compare the performance of MetaCluster 2.0 with existing binning algorithms, LikelyBin [22] and CompostBin [24] on clustering DNA fragments. The performance of MetaCluster 2.0 is better than these two algorithms in all datasets in [22]. We also compare the performance of MetaCluster with MEGAN on annotating DNA fragments. Experimental results show that MetaCluster annotates 20% more DNA fragments to the taxonomic tree correctly.

3.1 Clustering Performance

We randomly selected 300 species and downloaded their complete reference genomes from NCBI genomes database (ftp.ncbi.nih.gov/genomes/). These 300 species were selected among almost all the taxonomic groups with different ranks. In order to analyze the performance of MetaCluster under different situations, we generated 2,500 datasets from these genomics with different (1) taxonomic complexity (the number of species in the metagenomic dataset); (2) lengths of DNA fragments; (3) sequencing error rates and (4) relative abundance ratios (the ratio of DNA fragments among different species in the metagenomic dataset). We also compared the performance of MetaCluster, LikelyBin [22] and CompostBin [24] on five datasets provided in [22].

For each dataset, MetaCluster was used to cluster the DNA fragments. The clustering accuracy was calculated as the percentage of DNA fragments from the same species that are in the same cluster. Since our approach is unsupervised, no information about the species is needed to be given to MetaCluster including the exact number of species, where most

existing binning algorithms require the number of species as input parameters. It is because MetaCluster can merge several clusters together during the annotation process. However, in order to have a fair evaluation, the exact number of species in the dataset was given to the binning algorithms.

3.1.1 Taxonomic Complexity

When the DNA fragments come from related species, say species in the same Genus, the clustering process will be much difficult than those datasets with DNA fragments come from unrelated species, say species from different Orders. We divide the datasets into three categories: (1) DNA fragments from the same Family but different Genuses, (2) DNA fragments from the same Order but different Families, and (3) DNA fragments from different Orders. Table 1 summarises the experimental results of MetaCluster on these three categories when length-2000 error free DNA fragments are generated under the condition that the relative abundance ratio of each species in the datasets are the same. When the DNA fragments come from species in different Orders, MetaCluster performs well (about 90% accuracy) even there are 14 different species. When the DNA fragments come from species in the same Family or Orders, the performance of MetaCluster is still good especially when the number of species in the dataset is small.

Table 1. General performance based on different sample complexity and taxonomic differentia levels

Taxonomic Difference of Species	No. of species in datasets	Median	Lower Quartile	Upper Quartile
Genus	2	97.17%	94.69%	98.87%
	3	91.07%	81.41%	94.85%
Family	2	99.12%	95.23%	99.57%
	3	93.62%	87.90%	96.50%
Higher than Order	2	99.75%	98.32%	99.97%
	3	97.62%	94.96%	99.25%
	6	94.39%	92.80%	96.15%
	10	92.33%	88.02%	94.63%
	14	89.88%	80.03%	92.37%

3.1.2 Length of DNA Fragments

All the binning methods using *l*-mer distributions based on the observation that the *l*-mer distributions of DNA fragments from the same genome are more similar and consistent than DNA fragments from different genomes. So the stability of the *l*-mer frequency distribution is very essential for the binning performance. According to the studies in [18], the performance of binning is improved with longer DNA fragments. The performance improvement from longer fragments can be explained from two aspects. First, longer DNA fragments cover more the source genome, so can be considered more compositional similar and contain more information from the source genome. Second, longer DNA fragments provide more *l*-mers which provide statistically more reliable *l*-mer occurrence frequencies.

With the development of high-throughput sequencing, unlike the traditional Sanger sequencing which provides 2,000bp size reads, the new generation sequencing platform produces sequencing reads with length from about 400bp (454 FLX) to 120bp (Solexa) or even as short as 50bp (SOLiD). Although existing binning algorithms try to process the short reads, but until now, the extremely short reads, which range from 50 to 150bp are still out of reach of any non-alignment-based binning method. However, since the Spearman Footrule Distance definition is not hypersensitive to the variation of the absolute value of *l*-mer frequency, MetaCluster has a good performance even when the length of DNA fragments is 300bp. As shown in Figure 4, at the taxonomic differentia level of genus, MetaCluster achieves an average accuracy about 88%. Therefore, MetaCluster can be applied to binning the 454 pyro-sequencing reads.

When the length of DNA fragments increases from 500bp to 2000bp, there is an obvious improvement in accuracy. However, once the fragment is long enough, the accuracy improvement will taper off with further increase of fragment length. Therefore, MetaCluster performs well for the traditional Sanger sequencing reads.

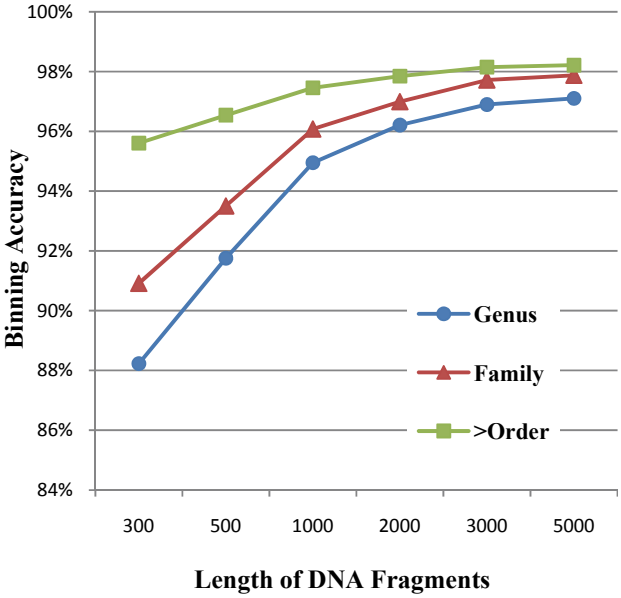

Figure 4. For each group of selected genomes, 6 datasets are generated with different fragment lengths (300bp, 500bp, 1000bp, 2000bp, 3000bp and 5000bp). With the increasing DNA fragment length, the average clustering accuracies based on three major taxonomic ranks tend to increase until a reasonable length is attained.

3.1.3 Sequencing Error Rates

Sequencing error is inevitable for metagenomics sequencing projects. Hence, error robustness is an important requirement for a successful binning algorithm. Although the typical sequencing error rate of existing commercial sequencing platforms is less than 2%, we generated test datasets with error rates ranging from 0% to

5% for checking the error robustness of MetaCluster. The performance of MetaCluster for different sequencing error rates is shown in Figure 5.

Even for a 5% error rate in the datasets, the accuracy of MetaCluster is only decreased by less than 1% when compared with error-free datasets. The error robustness property could be due to the chosen DNA composition features and the outlier filtering step. When compared with the alignment based algorithms where DNA substrings of 11bp are usually chosen as alignment seeds, MetaCluster uses 4-mer as the compositional feature. Any error on the read will affect 11 seeds for the alignment based algorithms and only four 4-mers for MetaCluster.

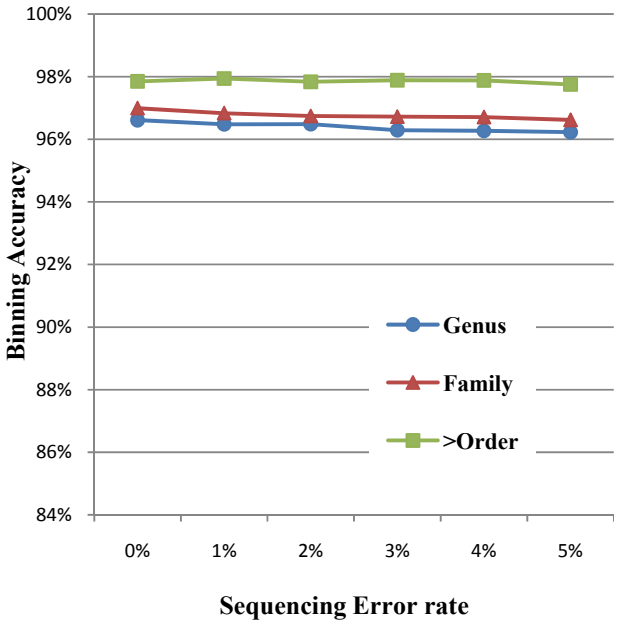

Figure 5. The three curves in the figure describe the binning average performance when the sequencing error increases from 0% to 5%.

3.1.4 Relative Abundance Ratio

Relative abundance ratio of species is a major factor affecting the performances of binning algorithms. When the relative abundance ratio between two species is high, existing binning algorithms cannot cluster DNA fragments from the same species because it is difficult to distinguish the DNA fragments in minority genome species from DNA fragments of rare patterns from heavily sampled species. We tested the performance of MetaCluster with abundance ratio of 1:1, 1:2, 1:3, 1:5 and 1:8, where the minority genome's DNA fragments take about 11.11% to 50% of the content. The variation trends are shown in Figure 6.

With the increase of abundance ratio, the decrease of Genus curve is obviously because the genomes of the species are too similar. However, when the DNA fragments come from species in different Families or Orders, the decreasing is not so significant. This result indicates that although MetaCluster works well for minority species, the performance for binning relatively rare species may still need further improvement.

3.1.5 Comparison with Existing Binning Algorithms

We compared the performances of MetaCluster, LikelyBin [22] and semi-supervised method CompostBin [24] based on the 5 genome pairs selected in the original paper of unsupervised binning tool LikelyBin. Each dataset contains two species with equal relative abundance ratio. There are 500 DNA fragments for each species in the datasets with fragment length as low as 400bp.

As the semi-supervised method, CompostBin requires several labelled DNA fragments as the initial seed for its clustering algorithm, different numbers of seeds are tested on performance. The binning performance of CompostBin is cited from the original paper of LikelyBin. In all five datasets, MetaCluster performs at least as good as all other algorithms. Besides, we also compared the approaches using other hundreds of datasets and in all test cases, the results are similar and in some cases, we can achieve about 15% higher accuracy. The details of this comparison will be given in the full paper.

The other highlight of MetaCluster is the computational efficiency. In practice, a 2-species dataset of 500 fragments per species, with length 800bp, consistently takes less than 10 seconds of CPU time to run on one Intel 3.0GHz core. For the same workload, LikelyBin takes about 360 to 450 seconds of CPU time.

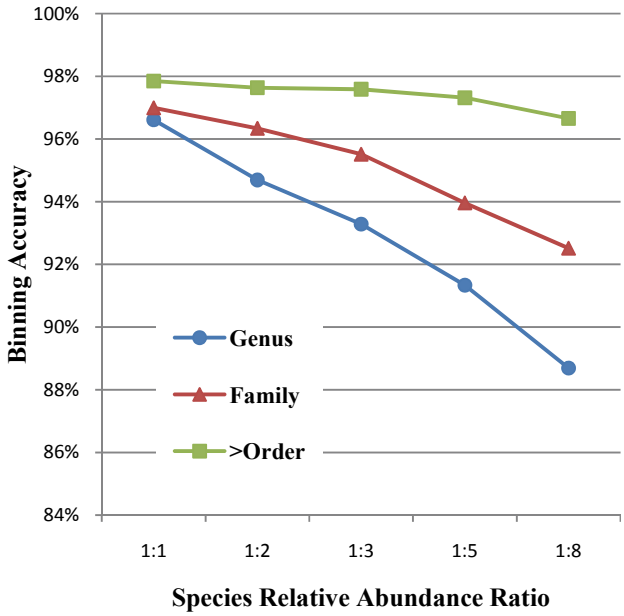

Figure 6. With the increasing of the relative abundance ratio of each level, the average clustering performances based on three major taxonomic ranges tend to decrease.

3.2 Accuracy of Annotating DNA fragments

In order to analyze the performance of MetaCluster on annotating DNA fragments to the taxonomic tree, first, all the 1140 complete bacterial chromosome genomes were downloaded from NCBI genomes database ftp server (ftp.ncbi.nih.gov/genomes/). The detailed taxonomic information was obtained from the NCBI taxonomic database (http://www.ncbi.nlm.nih.gov/taxonomy). The NCBI taxonomic tree structure could be downloaded from the NCBI taxonomy ftp server (ftp.ncbi.nih.gov/pub/taxonomy/).

Recall the precondition and methodology of our annotation process that a cluster of DNA fragments for taxonomic annotation are from an unknown genome which cannot be found in the NCBI database. It is highly probable that the genome sequence of some similar species in the level of genius and family or higher ranks may be available in the NCBI taxonomic tree. We can annotate the unknown genomes by estimating the similarity between fragments in the clusters and the reference genomes. To simulate the annotation task, in our experiment, we repeatedly selected one bacterial genome as the assumed unknown genome and this genome removed from the reference genome database. Meanwhile, 1,000 DNA substrings (fragments) are generated based on the selected genome to represent the cluster generated by the binning stage. In our experiment, the length of the DNA fragments is 2,000bp and sequencing error rate is 2%.

The results are shown in Figure 7. Although the annotating accuracy varies for different level between of the overlapping of distance between pairs of DNA fragments in different levels (Figure 3), the accuracy is over 87% in all levels.

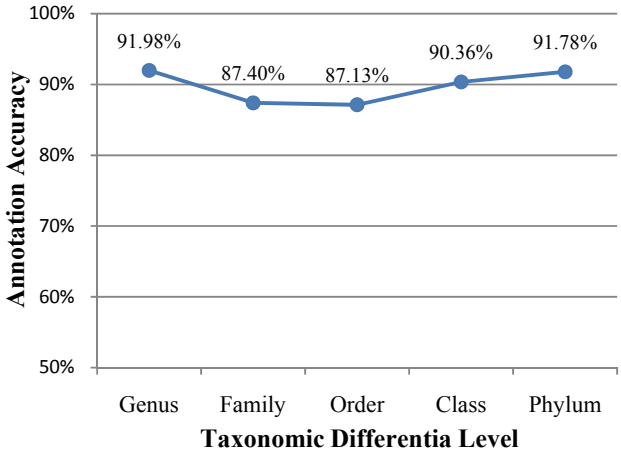

Figure 7. The annotation accuracy based on different taxonomic ranks. Both absolute and relative accuracy are illustrated here.

We also compared the performance of MetaCluster on annotation with MEGAN [8], a well applied binning tools based on the blast alignment result. MEGAN annotates reads by aligning reads to the reference genomes in the database. We randomly picked three genomes from the same Class but different Orders and removed these genomes from the NCBI taxonomic tree. For each genome, we generate 2,000 reads of length 2,000bp with error rate varies from 0 to 5%. MetaCluster and MEGAN were used to annotate these 6,000 reads. The clustering accuracy of MetaCluster varied from 92.5% to 95.3% and it also annotated the clusters to the correct taxonomic groups. The whole processes could be finished within 4 minutes based on one Intel 3.0GHz core. Based on Blastn alignment, in the same computer, MEGAN took over 120 minutes to finish annotation and about 1000 DNA fragments cannot be assigned to any reference genomes even we allow an e-value as large as 20. The clustering performance of MEGAN was only 78.78% (error free, e value = 2), 77.13 (5% error rate, e value = 2) and 76.63% (e value = 20 for much loose alignment during blast). Therefore, MetaCluster 2.0 out-performs MEGAN in both speed and accuracy.

4. DISCUSSION

In this paper, we have proposed MetaCluster 2.0 which could bin metagenomic sequencing datasets with high accuracy without any reference or background information and could also identify unknown genomes and annotate them with proper taxonomic labels. The methodology of our unsupervised binning step is based on the composition feature of *l*-mer distribution and Spearman Footrule Distance. The second annotation step is based on the plurality voting system with bottom-up integration strategy to annotate the clusters at particular taxonomic level.

For the unsupervised binning methods, the number of species in a metagenomic sequencing dataset is an essential input parameter that will affects the binning performance. The exact number of species in a MS dataset is usually unknown and is a complex research topic. However, to the best of our knowledge, all existing unsupervised and part of semi-supervised binning methods require the exact number of species be provided as input parameter.

In our research, the exact number of species inside the sample might not be needed. If the selected k value is less than the actual number of species inside the sample, the most similar species will be clustered together into some taxonomic specific groups. If the k value is larger than the actual number of species inside the sample, then some DNA fragments from the same genome will be divided into clusters. We tested different k values for some datasets and the result was satisfactory.

Although the exact value of k is not necessary, larger k value is always recommended because of two reasons. (1) Different clusters of the same genome can be annotated with the same taxonomic label. (2) When the abundance ratio between different species is large, a larger k will divide the majority species into many parts so as to maintain the balance among groups, therefore improve the clustering performance.

We have demonstrated this strategy with a simple case: randomly pick three genomes from the same Class but different Orders. For each genome, we generate 2,000 reads of length 2,000bp. The relative abundance ratio is 1:1:8. If k = 3, the clustering accuracy is only 83.45% due to the ambiguous cluster of minority species. If k = 5, the DNA fragments from majority species are divided into 3 clusters and the accuracy increases to 94.15% accordingly. After annotation, the clusters are annotated in the rank of Order. All the clusters are annotated correctly, and the clusters come from the same species are marked with same label thus the user can decide to merge them or not.

Acknowledgements

This research is supported by Project 30871393 of National Natural Science Foundation of China, Hong Kong GRF grant HKU 7117/09E and the HKU Genomics Strategic Research Theme Matching Fund.

5. APPENDIX

Algorithm 1: *K*-mean

Input:

S: set of input sequences

Output:

$C_1, C_2, \cdots, C_K$: partition of S

Procedures:

$E := 0$

Repeats M times

$(c'_1, c'_2, \cdots, c'_K) :=$ K randomly selected points from S

Repeats T times

for $i := 1$ to K

$C'_i := \{\}$

for each x in S

$i := \text{argmin}_{j=1}^{K} Distance_s(x, c'_j)$

$C'_i := C'_i \cup \{x\}$

for $i := 1$ to K

$c'_i := \frac{1}{|C'_i|} \sum_{x \in C'_i} x$

$E' := \sum_{i=1}^{K} \sum_{x \in C'_i} Distance_s{}^2(x, c'_i)$

if $E' < E$

$(C_1, C_2, \cdots, C_K) := (C'_1, C'_2, \cdots, C'_K)$

$E := E'$

Return $(C_1, C_2, \cdots, C_K)$

Algorithm 2: K-mean clustering with outliers filtering

Input:

S: set of input sequences

Output:

S': set of outliers

$C_1, C_2, \cdots, C_K$: partition of $S - S'$

Procedures:

$(C_1, C_2, \cdots, C_K) :=$ K-mean(S)

$S' := \{\}$

for i $:= 1$ to K

$c_i :=$ center of C_i

$\mu_i :=$ average distances, $\sum_{x \in C_i} Distance_s(x, c_i)/|C_i|$

$\sigma_i :=$ standard deviation of distances

$$\sqrt{\sum_{x \in C_i} (Distance_s(x, c_i) - \mu_i)^2 / (|C_i| - 1)}$$

for each $x \in C_i$

if $Distance_s(x, C_i) > \mu_i + 2\sigma_i$

$S' := S' \cup \{x\}$

$(C_1, C_2, \cdots, C_K) :=$ K-mean$(S - S')$

Return S' and $(C_1, C_2, \cdots, C_K)$

Algorithm 3: Annotation

Input:

S: set of input sequences

G: partition of all genome according to taxonomic level

Output:

S': set of outliers

$C_1, C_2, \cdots, C_K$: partition of $S - S'$

$\text{L}(C_1), L(C_2), \ldots, L(C_K)$: Classified label of each cluster

Procedures:

$(C_1, C_2, \cdots, C_k) :=$ Algorithm $2(S)$

for $i := 1$ to k

$w(g) := 0$ for all genome g

for each $x \in C_i$

$g^* :=$ nearest genome of x

$w(g^*) := w(g^*) + 1$

for $F \in G$

$w(F) := \sum_{g \in F} w(g)$

$\text{L}(C_i) :=$ argmax $_{F \in G} w(F)$

Return S', $(C_1, C_2, \cdots, C_K)$ and $(\text{L}(C_1), L(C_2), \ldots, L(C_K))$

6. REFERENCES

[1] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Jian, M., Zhou, Y., Li, Y., Zhang, X., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P. and Ehrlich, S.D. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59-65.

[2] Khachatryan, Z.A., Ktsoyan, Z.A., Manukyan, G.P., Kelly, D., Ghazaryan, K.A. and Aminov, R.I. (2008) Predominant role of host genetics in controlling the composition of gut microbiota. PLoS One 3, e3064.

[3] Amann, R.I., Binder, B.J., Olson, R.J., Chisholm, S.W., Devereux, R. and Stahl, D.A. (1990) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. Appl Environ Microbiol 56, 1919-25.

[4] Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H. and Smith, H.O. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66-74.

[5] Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S. and Banfield, J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37-43.

[6] Jones, B.V., Begley, M., Hill, C., Gahan, C.G. and Marchesi, J.R. (2008) Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. Proc Natl Acad Sci U S A 105, 13580-5.

[7] Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A.C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P. and Kyrpides, N.C. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods 4, 495-500.

[8] Huson, D.H., Auch, A.F., Qi, J. and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. Genome Res 17, 377-86.

[9] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402.

[10] Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., McGarrell, D.M., Garrity, G.M. and Tiedje, J.M. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res 33, D294-6.

[11] Chan, C.K., Hsu, A.L., Halgamuge, S.K. and Tang, S.L. (2008) Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics 9, 215.

[12] Garcia Martin, H., Ivanova, N., Kunin, V., Warnecke, F., Barry, K.W., McHardy, A.C., Yeates, C., He, S., Salamov, A.A., Szeto, E., Dalin, E., Putnam, N.H., Shapiro, H.J., Pangilinan, J.L., Rigoutsos, I., Kyrpides, N.C., Blackall, L.L., McMahon, K.D. and Hugenholtz, P. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol 24, 1263-9.

[13] Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C., Bork, P., Hugenholtz, P. and Rubin, E.M. (2005) Comparative metagenomics of microbial communities. Science 308, 554-7.

[14] Case, R.J., Boucher, Y., Dahllof, I., Holmstrom, C., Doolittle, W.F. and Kjelleberg, S. (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol 73, 278-88.

[15] Karlin, S. and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11, 283-90.

[16] Karlin, S. and Ladunga, I. (1994) Comparisons of eukaryotic genomic sequences. Proc Natl Acad Sci U S A 91, 12832-6.

[17] Karlin, S., Mrazek, J. and Campbell, A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179, 3899-913.

[18] Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. and Glockner, F.O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5, 163.

[19] Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. and Glockner, F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol 6, 938-47.

[20] Yang, B., Peng, Y., Leung, H., Yiu, S.M., Chen, J.C. and Chin, F. (2009) Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. In: DTMBIO '09: Proceeding of the third international workshop on Data and text mining in bioinformatics, pp. 3-10, ACM, Hong Kong, China.

[21] Diaconis, P. and Graham, R.L. (1977) Spearman's Footrule as a Measure of Disarray. Journal of the Royal Statistical Society. Series B (Methodological) 39, 262-268.

[22] Kislyuk, A., Bhatnagar, S., Dushoff, J. and Weitz, J.S. (2009) Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics 10, 316.

[23] Kendall, M.G. (1938) A new measure of rank correlation. Biometrika 30, 81-93.

[24] Chatterji, S., Yamazaki, I., Bai, Z.J. and Eisen, J.A. (2008) CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In: Research in Computational Molecular Biology, Proceedings, pp. 17-28.