

---

# A Robust and Accurate Binning Algorithm for Metagenomic Sequences with Arbitrary Species Abundance Ratio<sup>†</sup>

Henry C.M. Leung<sup>1</sup>, S.M. Yiu<sup>1</sup>, Bin Yang<sup>2</sup>, Yu Peng<sup>1</sup>, Yi Wang<sup>1</sup>, Zihua Liu<sup>1</sup>, Jingchi Chen<sup>2</sup>, Junjie Qin<sup>3</sup>, Ruiqiang Li<sup>3</sup>, Francis Y.L. Chin<sup>1\*</sup>

<sup>1</sup> Department of Computer Science, The University of Hong Kong, Hong Kong SAR, China

<sup>2</sup> State Key Laboratory of Bioelectronics, Southeast University, Nanjing, China

<sup>3</sup> BGI-Shenzhen, Shenzhen, China

---

## ABSTRACT

**Motivation:** With the rapid development of next-generation sequencing techniques, metagenomics, also known as environmental genomics, has emerged as an exciting research area which enables us to analyze the microbial environment in which we live. An important step for metagenomic data analysis is the identification and taxonomic characterization of DNA fragments (reads or contigs) resulting from sequencing a sample of mixed species. This step is usually referred to as “binning”. Binning algorithms that are based on sequence similarity and sequence composition markers rely heavily on the reference genomes of known microorganisms or phylogenetic markers. Due to the limited availability of reference genomes and the bias and low availability of markers, these algorithms may not be applicable in all cases. Unsupervised binning algorithms which can handle fragments from unknown species provide an alternative approach. However, existing unsupervised binning algorithms only work on datasets either with balanced species abundance ratios or rather different abundance ratios, but not both.

**Results:** In this paper, we present MetaCluster 3.0, an integrated binning method based on the unsupervised top-down separation and bottom-up merging strategy, which can bin metagenomic fragments of species with very balanced abundance ratios (say 1:1) to very different abundance ratios (e.g. 1:24) with consistently higher accuracy than existing methods.

**Availability:** MetaCluster 3.0 can be downloaded at <http://i.cs.hku.hk/~alse/MetaCluster/>

## 1 INTRODUCTION

Traditional microbial genomic studies usually focus on one single individual bacterial strain due to experimental limitations. In fact, all microorganisms in a habitat have various functional effects on one another and their hosts. For example, the diversity of microbes in humans is shown to be associated with common diseases such as Inflammatory Bowel Disease (IBD) (Qin et al., 2010) and gastrointestinal disturbance (Khachatryan et al., 2008). Genomic analysis on the collective genomes of all microorganisms from an environmental sample (also known as metagenomics, environmental genomics, or community genomics) becomes essential. One major difficulty of metagenomics lies in the fact that most bacteria (up to 99%) found in environmental samples are unknown and cannot be

cultivated and isolated under laboratory conditions (Amann et al., 1990). With high-throughput sequencing technology, one possible solution is to directly sequence the DNA fragments of multiple species obtained from the mixed environmental DNA sample (Venter et al., 2004). Some well-known metagenomics projects, including the Acid Mine Drainage Biofilm (AMD) project which analyzes dozens of species (Tyson et al., 2004) and the recent Human Gut Microbiome (HGM) project which involves thousands of species, (Jones et al., 2008) study fragments obtained from this sequencing approach.

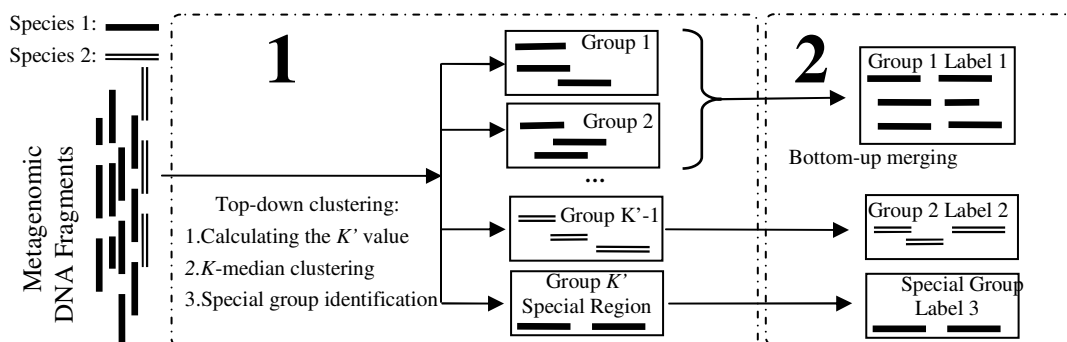
DNA fragments of a metagenomics project are usually from multiple genomes and most of the genome sequences are unknown. An important step in metagenomic analysis is to group DNA fragments from similar species together (referred to as *binning*) (Mavromatis et al., 2007) to obtain the microbe distribution of the sample and identify species (including unknown species) within the sample. Depending on different research needs, the binning process could be performed on different taxonomic levels from Kingdom (the highest level) to Species (the lowest level).

Traditional binning methods can be roughly classified as similarity-based and composition-based. Similarity-based methods (Huson, et al., 2007) align each DNA fragment to known reference genomes. Based on the alignment results (e.g. BLAST hits or selected phylogenetic specific marker genes (Altschul, et al., 1997)), each fragment is assigned to the taxonomic class of the similar reference genomes. Similarity-based methods are usually limited by the availability of known microorganism genomes given that less than 1% of microorganisms have been cultured and sequenced. On the other hand, composition-based methods group DNA fragments in a supervised or semi-supervised manner using generic features such as genome structure or composition. Structural features, such as composition features of reference genomes or taxonomic marker regions (e.g. 16S rRNA (Cole et al., 2005), *recA* and *rpoB* are commonly accepted fingerprint genes), are extracted and used to construct classifiers (Chan et al., 2008; Chatterji et al., 2008) for determining DNA fragments from different species or constraints for semi-supervised clustering. These composition-based methods usually suffer from the low availability and reliability of taxonomic markers. For example, studies on the enhanced biological phosphorus removing (EBPR) sludge (Garcia Martin et al., 2006), Sargasso Sea (Venter et al., 2004) and the Minnesota soil samples (Tringe et al., 2005) indicate that only 0.17%, 0.06% and 0.017% of the DNA fragments respectively are known to carry 16S rRNA markers, and the figures are still less than 1% even if

---

<sup>†</sup> This paper is partial supported by the GRF grant HKU 719709E, HKU 711611 and HKU SPACE Research Fund.

\* To whom correspondence should be addressed.



**Figure 1.** The pipeline of MetaCluster 3.0 is divided into two major phases: Top-down clustering and Bottom-up merging.

more markers such as *recA* and *rpoB* are included. Moreover, the reliability of taxonomic markers has also been challenged (Case et al., 2007) as some species may share multiple markers with other species or multiple kinds of 16S rRNA molecules exist in a single bacterium due to the high mutation and gene exchange ratio of microbes, which might lead to incorrect classifications.

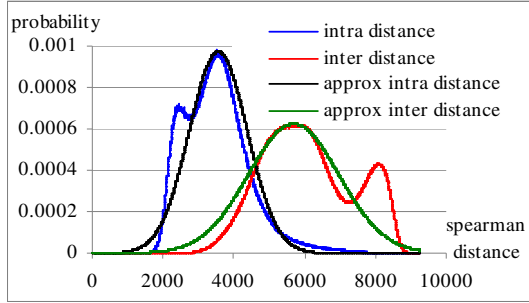
A more promising direction is to use unsupervised binning algorithms based on the occurrence frequencies of  $l$ -mers (short DNA substrings of length  $l$ ) of the DNA fragments (Karlin and Burge, 1995; Karlin and Ladunga, 1994). In these approaches, each fragment can be regarded as a vector containing the occurrence frequencies of all possible  $l$ -mers in the fragment. The rationale behind these approaches is based on the observation that the  $l$ -mer distributions of the fragments in the same genome are more similar than those  $l$ -mer distributions of two unrelated species (Chor et al., 2009; Karlin et al., 1997; Yang et al., 2010). Based on this feature, many algorithms (e.g. TETRA (Teeling et al., 2004), MetaCluster (Yang et al., 2009), MetaCluster 2.0 (Yang et al., 2010) and LikelyBin (Kislyuk et al., 2009)) have applied different unsupervised clustering methods on fragments based on the  $l$ -mer distributions with great success when the abundance ratios of the species in the sample are almost the same. These algorithms tend to cluster fragments into clusters with similar size, and the binning performance of these algorithms will significantly be degraded when the abundance ratios of species are very different, e.g. 1:8 or 1:10. To tackle this problem, another unsupervised clustering method called AbundanceBin (Wu and Ye, 2010) has been introduced recently which models the distribution of fragments from each species by different Poisson distributions. Fragments from species with high abundance ratios are usually sampled more while fragments from species with low abundance ratios are usually sampled less. Thus, fragments from species with different abundance ratios can be separated by their sampling rates. However, when the species have similar abundance ratios, say 1:1 or 1:2, the sampling rates of fragments from different species are similar and AbundanceBin fails to separate the fragments from these species. AbundanceBin also considers the case when the number of species  $k$  is unknown and determines  $k$  automatically. The estimation method seems not very effective and cannot predict the correct value when  $k > 3$  in our experiments. None of the above algorithms can handle datasets from balanced species abundance ratios to very different species abundance ratios. In real situations, the abundance ratios are usually unknown and it is desirable to have an algorithm handles datasets with arbitrary abundance ratios.

AbundanceBin takes advantage of the differences in abundance ratio of the species to separate the fragments and it is not obvious how to extend AbundanceBin to handle datasets with very balanced abundance ratio. On the other hand, approaches that handle datasets with balanced abundance ratio usually try to cluster fragments into  $k$  equal-sized clusters by assuming the number of species in the dataset is  $k$ . Thus, species with low abundance ratio are forced to group wrongly with other species.

In this paper, we present MetaCluster 3.0, an integrated binning algorithm based on two phases: top-down clustering (Phase 1) and bottom-up merging (Phase 2). In Phase 1, we first separate fragments into small groups (clusters) with similar sizes and try to guarantee that the majority of the fragments belong to the same species. After Phase 1, it is possible that fragments from the same species are grouped into different clusters, so in Phase 2, we try to combine these clusters together. Figure 1 outlines the two phases of MetaCluster 3.0. This 2-phase strategy relies on the following key observations. First, the difference (we capture this difference using the Spearman distance measure) between two  $l$ -mer distributions of fragments from the same species follows a normal distribution. The same is true for the two  $l$ -mer distributions of fragments from species of different families. This observation is supported by an empirical study (see Section 2.2). Second, the differences in the two normal distributions allow us to derive a probabilistic model to determine how many clusters ( $k'$ ) we should use in order to guarantee that most of the fragments in the same cluster belong to the same species in Phase 1 and when to merge the clusters to make sure that there will not be too many fragments from other species being merged together.

Based on our probabilistic model, if the abundance ratios of the species are more or less the same, the value of  $k'$  will be similar to the number of species. If the species have very different abundance ratios,  $k'$  tends to be large and the clusters will be small so that fragments from minority species could be assigned to separate clusters without mixing with fragments from other species. Although fragments of majority species may then be put in different clusters, Phase 2 will try to merge them back. Thus, the issue of varying species abundance ratio is handled.

To summarize, MetaCluster 3.0 can (1) determine automatically the number of different species in the sample, which is required as an input parameter for most unsupervised algorithms (Kislyuk et al., 2009; Teeling et al., 2004; Yang et al., 2009; Yang et al., 2010), and (2) classify accurately the metagenomic fragments with balanced species abundance ratios, which cannot be handled by AbundanceBin (Wu and Ye, 2010), to very different species abun-



**Figure 2.** Probability density functions of the Spearman distance between two fragments from the same species (intra distance) and between two fragments from the same order but different families (inter distance). Approx intra distance and approx inter distance is the normal distribution approximation of the two distances.

dance ratio, which cannot be handled by other unsupervised algorithms (Kislyuk et al., 2009; Teeling et al., 2004; Yang et al., 2009; Yang et al., 2010) and combinations of these situations, say 1:3:3:9, which cannot be handled by any unsupervised algorithms.

## 2 METHODS

In this section, we first define the  $l$ -mer feature vector of a fragment that captures the  $l$ -mer frequency distribution of the fragment. Then, we describe the Spearman Footrule distance (Diaconis and Graham, 1977) to represent the difference (distance) between two  $l$ -mer feature vectors or their corresponding fragments. Besides Spearman Footrule distance, we have also tried different similarity measures such as Kendall’s Tau (Kendall, 1938) and those mentioned in (Wu and Ye, 2010; Yang et al., 2009). Spearman Footrule distance seems to be better in terms of performance in our clustering and computational complexity. We remark that there may also be other measures which are appropriate to solve the problem. Next, we will show the result of an empirical study which confirms our key observations. Lastly, we describe the details of top-down clustering (Phase 1) and bottom-up merging (Phase 2) together with our probabilistic model which is used to determine the number of clusters to be used in Phase 1 and when to merge two clusters in Phase 2.

### 2.1 $l$ -mer frequency and distance definition

The *DNA composition features* of each DNA fragment are represented by the  $l$ -mer frequencies of the DNA fragment. As there are 4 different DNA nucleotides, there are at most  $4^l$  kinds of  $l$ -mers in a DNA sequence. If a sliding window of width  $l$  is slid along each DNA fragment of length  $n$  and the frequency of every  $l$ -mer, say  $f_w$ ,  $w \in \{A,C,G,T\}^l$ , is recorded, the total number of  $l$ -mers in a DNA fragment would be  $\sum f_w = n - l + 1$ . For example, a DNA fragment of length 500 nt has 497 4-mers. The *DNA feature vector* is defined as  $[f_1, f_2, \dots, f_{N(l)}]$ , where  $N(l)$  is the number of different  $l$ -mers. 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 together and this process will reduce the size of vector by half, i.e.  $N(l) = 4^l/2$ , if  $l$  is odd;  $N(l) = (4^l + 4^{l/2})/2$ , if  $l$  is even.

As mentioned in (Chor et al., 2009; Zhou et al., 2008), setting  $l = 4$  is the best (among  $l = 2$  to 7) when barcoding a genome with DNA fragment size from 1,000 nt to 10,000 nt. Each DNA fragment will be represented by a feature vector with 136 components

and the input metagenomic sequencing dataset can be transformed to an  $n \times 136$  matrix with  $n$  rows representing  $n$  DNA fragments. Recall that our binning method is based on the observation (Chor et al., 2009; Teeling et al., 2004) that the  $l$ -mer distributions of those DNA substrings (fragments) from the same genome are similar. The similarity of 4-mer distribution is not limited to the coding region but the whole genome sequence (Chor et al., 2009; Zhou et al., 2008). We compute the difference of two  $l$ -mer distributions from two fragments by measuring the Spearman Footrule distance between their corresponding  $l$ -mer feature vectors.

Spearman Footrule distance (henceforth referred as *Spearman distance*) is defined as follows. Consider two DNA fragments A and B with the following 4-mer feature vectors A:  $(a_1, a_2, \dots, a_i, \dots, a_j, \dots, a_{N(l)})$  and B:  $(b_1, b_2, \dots, b_i, \dots, b_j, \dots, b_{N(l)})$ . The Spearman distance is based on an intuitive definition for comparing two ordered lists. Let  $r^A(a_i)$  be the rank of  $a_i$  in the sorted list of  $a_i$ ’s and  $r^A(b_i)$  be the rank of  $b_i$  in the sorted list of  $b_i$ ’s. Then the Spearman distance is defined as  $dist_s(A,B) = \sum |r^A(a_i) - r^B(b_i)|$ . The smaller the value of the metric, the more similar the vectors are. For vectors with size  $k$ , the distance value can range between 0 and  $k(k+1)$ . Compared with other distance metrics that rely on the exact value of each entry in the feature vectors, Spearman distance, which relies on the rank of the entries, is less sensitive to those entries with unexpectedly large values. Moreover, the Spearman distance gives a more global view of the distance of two feature vectors with respected to all the entries.

### 2.2 Spearman distance distribution

To confirm our observation that both the Spearman distance distributions of the differences between two  $l$ -mer distributions of fragments (pairwise fragment distances) from the same species and those from species of different families can be approximated by a normal distribution, we conduct an empirical study for 1,000 genomes. For each genome, we randomly select 1 million pairs of fragments of 1,000 nt long, and compute the Spearman distances of all pairs. This distance distribution is referred as *intra-distance distribution* (see Figure 2). For fragments from different families, we select 10,000 pairs of genomes in which the two genomes of each pair belong to different families but are of the same order. For each pair of genomes, we select one fragment of length 1,000 nt from each genome and compute the Spearman distance of these two fragments. We repeat this randomly for 1 million pairs of fragments. This distance distribution is referred as the *inter-distance distribution* (see Figure 2). From our empirical study, we can see that these two distributions can be approximated by normal distributions and there is a significant difference between these two distributions. In fact, the distribution can be modeled by a mixed Gaussian distribution because of differences in inter and intra distances among different genomes. However, as we assume that there is no information of what kinds of species are in the mixture, we used normal distribution for approximation only. In the following, we describe the details of the two phases (top-down clustering and bottom-up merging) and how we make use of the difference in the intra-distance and inter-distance distributions to guarantee the accuracy of these two phases in MetaCluster 3.0.

### 2.3 Top-down clustering

In this step, we apply the simple  $k$ -median algorithm<sup>1</sup>(Jain and Dubes, 1981) to cluster the fragments into  $k'$  clusters of similar sizes.  $k$ -median algorithm repeatedly assigns feature vector to the closest cluster and select a feature vector in each cluster as the center with the following objective function

$$MinE = \sum_{i=1}^{k'} \sum_{A \in C_i} dist_s(A, c_i)$$

where feature vector  $c_i$  is the center of cluster  $C_i$  and  $dist_s(A, c_i)$  is the spearman distance between feature vectors  $A$  and  $c_i$ .

In MetaCluster 3.0, the value of  $k'$  is determined automatically based on a probabilistic model by restricting the expected number of false positive fragments (from other species) in a cluster to be limited by some predefined threshold  $t \times$  size of the cluster,  $t \in (0,1]$ . The details of how to determine  $k'$  are given below. Since the  $k$ -median algorithm is a greedy algorithm, it is repeated several dozen times with different initial clustering centers. The one that gives the minimum objective function value will be selected.

Now, we show how to determine  $k'$ . By dividing  $n$  fragments into  $k'$  clusters, the average cluster size is  $n/k'$ . In each cluster, there are two sets of fragments, fragments from the same species as the center and fragments from species different from the center. The distance between each fragment and the center from the same species can be approximated by  $N(\mu_{intra}, \sigma_{intra}^2)$  while the distance between each fragment and the center from different species can be approximated by  $N(\mu_{inter}, \sigma_{inter}^2)$  (Figure 2). Given a cluster  $C_i$  with the total distance between the center  $c_i$  and each feature vector in the cluster  $\sum_{A \in C_i} dist_s(A, c_i)$  equals a particular value  $d_i$ . If  $s$  out of  $n/k'$  fragments (including the center) in the cluster are sampled from the same species with average distance (intra-species distance) between the center and the rest  $s-1$  fragments be  $x$ , the probability that there are  $n/k' - s$  false positives equals the probability that the average distance (inter-species distance) between the center and the  $n/k' - s$  fragments from different species be  $(d_i - (s-1)x)/(n/k' - s)$ , which follows the Gaussian distribution  $N(\mu_{inter}, \sigma_{inter}^2 / (n/k' - s))$ . By considering all possible values of  $s$  and  $x$ , the expected number of false positives in a cluster can be calculated as follows

$$\sum_{s=2}^{n/k'} s \int_0^{\infty} f_{\mu_{inter}, \sigma_{inter}^2 / (s-1)}(x) \left[ \int_0^{(d_i - (s-1)x)/(n/k' - s)} f_{\mu_{intra}, \sigma_{intra}^2 / (n/k' - s)}(y) dy \right] dx$$

where  $f_{\mu, \sigma^2}(x) = e^{-(x-\mu)^2 / (2\sigma^2)} / \sqrt{2\pi\sigma^2}$  is the probability density function of a normal distribution with mean  $\mu$  and variance  $\sigma^2$ .

Since the expected number of false positives decreases with the value of  $k'$ , MetaCluster 3.0 will increase the value of  $k'$  until the expected number of false positives in a cluster  $\leq tn/k'$ . In the experiments, we set  $t = 5\%$  such that the expected accuracy is over 95% for the first phase.

Based on the above calculation, we expect that  $k'$  can be much larger than the number of species if the species have very different abundance ratios such that fragments from species with high abundance ratios will be divided into more clusters while fragments from species with low abundance ratios will be grouped into a single cluster or fewer clusters.

<sup>1</sup> We use  $k$ -median clustering algorithm as it is easy to compute. Further investigation on the effectiveness of different clustering algorithms should be conducted.

As for the same genome, the  $l$ -mer distribution of some special genome region (such as insertion and exogenous transferred regions) can be very different from general genome regions. These data points could be considered outliers and should be removed. In MetaCluster 3.0, those data points with center distance larger than  $\mu + 2\sigma$  should be removed as outliers, where  $\mu$  and  $\sigma$  are the average distance and standard deviation between a data point in the cluster and the center respectively. In some cases, the number of outlier DNA fragments from the majority species could be very large and might have special biological meaning. So these fragments will be grouped together as some special clusters which will be excluded from the merging phase, but reported specifically for the attention of biologists.

## 2.4 Bottom-up merging of the clusters

After dividing the DNA fragments into  $k'$  clusters, a bottom-up merging phase is introduced to merge the clusters from the same species into one cluster based on the inter-cluster similarity, i.e. inter-cluster distance. The inter-cluster distance of cluster  $C_1$  and cluster  $C_2$  is taken to be the average of all distances between pairs of DNA fragments  $A$  in  $C_1$  and  $B$  in  $C_2$ .

$$dist(C_1, C_2) = \frac{\sum_{A \in C_1} \sum_{B \in C_2} dist_s(A, B)}{|C_1| \cdot |C_2|}$$

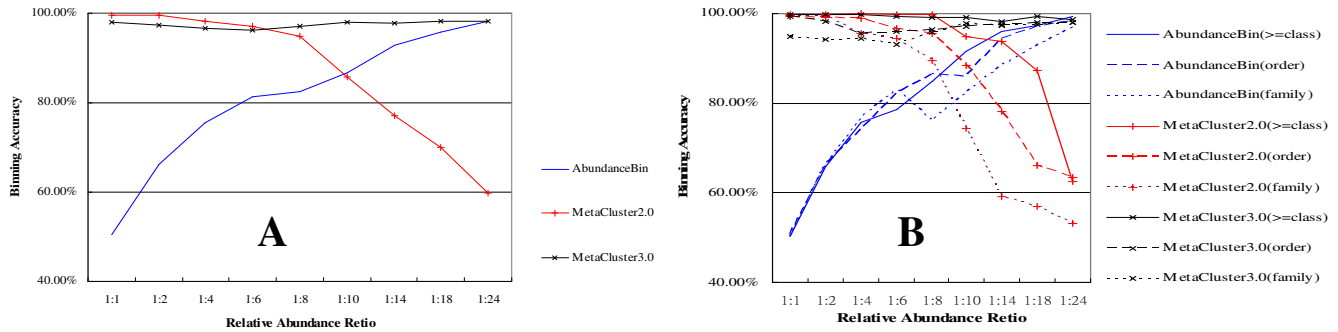
When the number of species  $k$  in the sample is known, MetaCluster 3.0 merges the pair of clusters with the minimum inter-cluster distance greedily until there are  $k$  large clusters. In practical situations, the number of species  $k$  is usually unknown and MetaCluster 3.0 should determine when to stop merging so that clusters that contain fragments from different species will not be merged into a cluster. Based on the observation that the Spearman distance between two fragments from the same species is usually smaller than the Spearman distance between two fragments from different species, MetaCluster 3.0 merges two clusters  $C_1$  and  $C_2$  with average intra-cluster distance  $d_1$  and  $d_2$  respectively if and only if the inter-cluster distance  $dist(C_1, C_2)$  is similar to  $d_1$  and  $d_2$ , i.e.  $\alpha \cdot dist(C_1, C_2) \leq average(d_1, d_2)$  for some threshold  $\alpha \in (0,1]$ . The value of threshold  $\alpha$  can be determined by minimizing the expected false negative and false positive fragments. Assume all fragments in  $C_1(C_2)$  are sampled from the same species, the intra-cluster distance can be modeled by the intra-species distance distribution. The probability that MetaCluster 3.0 does not merge two clusters incorrectly (false negative) can be calculated as follow:

$$\begin{aligned} & P(\text{false negative}) \\ &= \int_0^{\infty} P(\text{average}(d_1, d_2) = x) P(\text{dist}(C_1, C_2) > x / \alpha | C_1, C_2 \text{ from same species}) dx \\ &= \int_0^{\infty} f_{\mu_{intra}, \sigma_{intra}^2}(x) \int_{x/\alpha}^{\infty} f_{\mu_{intra}, \sigma_{intra}^2}(y) dy dx \end{aligned} \quad (1)$$

Similarly, the probability that MetaCluster 3.0 merges two clusters incorrectly (false positive) can be calculated.

$$P(\text{false positive}) = \int_0^{\infty} f_{\mu_{intra}, \sigma_{intra}^2}(x) \int_0^{x/\alpha} f_{\mu_{inter}, \sigma_{inter}^2}(y) dy dx \quad (2)$$

For  $\mu_{intra} = 3550$ ,  $\sigma_{intra} = 820$ ,  $\mu_{inter} = 5676$  and  $\sigma_{inter} = 1278$  estimated from bacteria genome, setting the threshold  $\alpha = 0.79$  can minimize the expected false negative and false positive (1) + (2) fragments. This threshold is similar to the optimal threshold  $\alpha = 0.83$  found in the simulated data. Unlike all other unsupervised



**Figure 3.** Comparison of accuracy among MetaCluster 3.0, MetaCluster 2.0 and AbundanceBin. (A) Overall performance of all datasets. (B) Performance for Class, Order and Family datasets.

approaches which do not provide any taxonomic annotation for the clusters, MetaCluster 3.0 can label (annotate) the clusters with taxonomic information by calculating the average Spearman distance between each cluster and the 4-mer feature vectors of the known genome. Although many genomes are still unknown, it will provide an approximated annotation at high taxonomic ranks such as Family or Order, which helps the biologists to determine follow-up experiments for further investigation.

The time complexity of MetaCluster3.0 is  $O(n^2)$  where  $n$  is the number of fragments. MetaCluster 3.0 does not use a lot of memory. Detailed information about the actual running time and the amount of memory used will be given in the next section.

### 3 RESULTS

In this section, we analyze the performance of the binning algorithm, MetaCluster 3.0, based on the simulated metagenomic datasets. We compare the performance of MetaCluster 3.0 with AbundanceBin (Wu and Ye, 2010) and our previous version MetaCluster 2.0 (Yang et al., 2009). We have not compared other unsupervised binning algorithms because MetaCluster 2.0 outperforms these algorithms in similar experimental setting (Yang et al., 2009). We use the default parameters for AbundanceBin and MetaCluster 2.0. We have conducted three sets of experiments. (1) We fix the number of species to be 2 and vary the abundance ratio from the balanced situation 1:1 to the unbalanced situation of 1:24. We assume that the number of species in the dataset is known. The performance of our new version MetaCluster is consistently more accurate for all datasets with different abundance ratios. (2) We also compare the performance of MetaCluster 3.0 with AbundanceBin based on datasets with more species with different abundance ratios. In this set of experiments, we also assume that the number of species is known. The results show that MetaCluster 3.0 outperforms AbundanceBin. In particular, the accuracy of MetaCluster 3.0 is 3 times better than that of AbundanceBin when the species abundance ratio is balanced. (3) Lastly, we demonstrate that MetaCluster 3.0 works better than AbundanceBin if the number of species in the dataset is unknown. In all the experiments, we use the parameters  $t = 5\%$  and  $\alpha = 0.8$  for MetaCluster 3.0. We have varied the values of these parameters and the results are similar.

#### 3.1 The datasets

A total of 120 pairs (240 genomes) of bacteria are randomly selected and their complete reference genomes are downloaded from the NCBI reference genomes database: <ftp.ncbi.nih.gov/genomes/>

to generate 1080 test datasets. These 120 pairs of genomes are equally divided into 3 major testing categories according to different taxonomic differential levels. (1) Family: DNA fragments from the same Order but different Families, (2) Order: DNA fragments from the same Class but different Orders, and (3)  $\geq$  Class: DNA fragments from different Classes. For each pair of bacteria genomes, 9 synthetic metagenomic datasets are generated with sequencing fragments sampled from these two genomes of different relative abundance ratios, as 1:1, 1:2, 1:4, 1:6, 1:8, 1:10, 1:14, 1:18 and 1:24. The lengths of these DNA fragments is 1,000 nt and the sequencing error rate is 1% which is the sequencing error rate of major next generation sequencing platform, i.e. Roche 454, Illumina Solexa and Applied Biosystems SOLiD. This dataset is used to estimate the performance of MetaCluster 3.0 under different species similarity. Another set of 40 multi-species test datasets containing 3 and 4 genomes are also generated similarly. The details of these datasets and the binning performance will be discussed in the following.

For each dataset, MetaCluster 2.0, MetaCluster 3.0 and AbundanceBin were used to cluster the DNA fragments. Each cluster was assigned to the species with the largest number of fragments in the cluster. The binning accuracy, following other approaches (Wu and Ye, 2010; Yang et al., 2009; Yang et al., 2010), is defined as the total number of true positive fragments divided by the total number of fragments in the clusters. The performance of the algorithms is measured by the average accuracy among all datasets. Since our approach is unsupervised, no information about the species is needed to be given to MetaCluster, while most existing binning algorithms require the number of species as input parameter. However, in order to have a fair evaluation and comparison, the exact number of species in the dataset was given to the binning algorithms including MetaCluster 3.0 for the experimental results described in Sections 3.2 and 3.3. The performance of MetaCluster 3.0 on determining the number of species  $k$  in the sample will be shown in Section 3.4, whereas in these experiments, the value of  $k$  is not given to the tools.

#### 3.2 Experiments on different abundance ratio

For all the unsupervised binning methods, relative abundance ratio of species is a major factor affecting the performance of binning algorithms. We first use the typical datasets with two species for evaluation. We compare the performance of MetaCluster 3.0 with AbundanceBin and MetaCluster 2.0 using abundance ratio of 1:1, 1:2, 1:4, 1:6, 1:8, 1:10, 1:14, 1:18 and 1:24, where the minority genome's DNA fragments are about 50% to only 4% of the total

**Table 1.** The binning performance of MetaCluster 3.0 and AbundanceBin based on the multi-species test datasets.

3 Species	1:1:1			1:3:9		
	Avg	Max	Min	Avg	Max	Min
MetaCluster 3.0	97.36%	99.97%	83.24%	95.89%	99.96%	87.52%
AbundanceBin	33.63%	35.53%	33.33%	81.44%	91.30%	69.23%
4 Species	1:1:1:1			1:3:3:9		
	Avg	Max	Min	Avg	Max	Min
MetaCluster 3.0	97.12%	99.80%	77.23%	90.49%	99.90%	72.55%
AbundanceBin	25.4%	26.71%	25.00%	69.39%	75.04%	56.25%

**Table 2.** The predicted number of species in a sample by MetaCluster 3.0 and AbundanceBin based on the multi-species test datasets. The average accuracy is the overall performance of the algorithms and the number of clusters is the number of times the algorithms estimate the number of clusters in the dataset as the corresponding number.

3 Species	1:1:1					1:3:9							
	Avg Acc.	Number of clusters				Avg Acc.	Number of clusters						
		1	2	3	4		1	2	3	4			
MetaCluster 3.0	95.53%	-	10%	90%	-	97.78%	-	5%	95%	-			
AbundanceBin	33%	100%	-	-	-	68.55%	30%	70%	-	-			
4 Species	1:1:1:1					1:3:3:9							
	Avg Acc.	Number of clusters					Avg Acc.	Number of clusters					
		1	2	3	4	5		1	2	3	4	5	6
MetaCluster 3.0	96.74	-	-	-	95%	5%	96.45	5%	-	-	50%	35%	10%
AbundanceBin	25%	100%	-	-	-	-	67.68%	25%	75%	-	-	-	-

content. The performances of these three methods are shown in Figure 3A. The performance of MetaCluster 2.0 significantly deteriorates when the abundance ratio is larger than 1:8. On the other hand, the performance of AbundanceBin is very good when abundance ratio is larger than 1:10, but not acceptable when the abundance ratio is less than 1:8. In contrast with these two methods, the binning accuracy of MetaCluster 3.0 is consistently high for complex species abundance ratio from exactly equal situation, say 1:1, to the extremely unbalance situation, say 1:24.

We also compare the performance of these three methods based on three different taxonomic levels described in Section 3.1, i.e., (1) Family, (2) Order, and (3) Class. Figure 3B shows the binning accuracy of MetaCluster 3.0, MetaCluster 2.0 and AbundanceBin based on these three testing categories. MetaCluster 3.0 performs consistently well (over 90% accuracy) even the taxonomic level is as low as Family.

### 3.3 Experiments on multi-species test datasets

Increasing the number of species in the datasets will significantly increase the difficulty for binning algorithms. With more species, the species abundance distribution becomes more complicated. The most common situation is when the abundance ratio between the majority species and minority species is very large with other species having more or less the same abundance ratios. We generate 2 categories of multi-species test datasets. One category contains 20 3-species combinations, where for each combination, we generate one dataset of balanced abundance ratio say, 1:1:1 and one dataset of abundance ratio is 1:3:9. The other category contains 20 4-species combinations. For each combination, we generate one dataset of balanced abundance ratio say, 1:1:1:1 and one dataset of abundance ratio is 1:3:3:9. These latter unbalanced datasets could be considered as the simplified simulated model for practical multi-species cases with arbitrary abundance ratios.

The binning performances of MetaCluster 3.0 and AbundanceBin are shown in Table 1. AbundanceBin performs much better for the case when the species have quite different abundance ratios, but the accuracy drops substantially when the species have the same abundance ratios. This result matches our observation and indicates that the AbundanceBin relies heavily on the very unbalanced abundance ratio among species to do the clustering. On the other hand, the performance of MetaCluster 3.0 is quite consistent in both cases although there is a drop in the accuracy for the unbalanced case. It seems that MetaCluster 3.0 may be more suitable for practical applications.

### 3.4 Unknown number of species

We compare the performance of MetaCluster 3.0 and AbundanceBin when the numbers of species in the samples are unknown. Recall that we set a threshold  $\alpha = 0.8$  for determining the stopping condition in the merging phase in MetaCluster 3.0. The average accuracy of the resulting clusters can be found in Table 2 and are consistent with all other experiments. MetaCluster 3.0 performs better than AbundanceBin, in particular for the case of balanced species abundance ratios. Note that the accuracy of MetaCluster 3.0 drops about 2% when  $k$  is unknown. However, the accuracy of AbundanceBin drops 10% for the datasets of 3 species with abundance ratio 1:3:9 and 2% for other datasets.

In addition to the average accuracy of the resulting clusters, we also compare the number of clusters reported by the tools. For each test case of fixed number of species and abundance ratios, we have repeated the experiments 20 times. In Table 2, we record the number of clusters reported in each case and the percentage of cases that the tools report correctly. Note that if the case has 3 species, the perfect answer should be 3 resulting clusters. In general, MetaCluster 3.0 can predict the number of species in the datasets more accurately than AbundanceBin which usually underestimates the



**Table 3.** The running time and memory consumption of MetaCluster 3.0, MetaCluster 2.0 and AbundanceBin for different data set size.

	Running Time			Memory consumption		
	10,000 fragments	50,000 fragments	100,000 fragments	10,000 fragments	50,000 fragments	100,000 fragments
MetaCluster 3.0	15 s	5 min	17 min	173 M	354 M	581 M
MetaCluster 2.0	14 s	5 min 18 s	20 min	175 M	356 M	583 M
AbundanceBin	2 min 21 s	18 min	37 min	683 M	1.40 G	1.98 G

number of species in a sample. For the dataset of 4 species with abundance ratio 1:3:3:9, MetaCluster 3.0 performs not as well with 35% of the dataset clustered into 5 groups, instead of 4. It is because MetaCluster 3.0 separates the reads in the species with the highest abundance ratio into two parts, one with the normal fragments and the other with fragments from special region, say the coding regions, which have a different  $k$ -mer distribution as other regions in the genome. It may be solved by lowering the threshold  $\alpha$  for those data with extremely unbalanced abundance ratios. However, lowering the threshold  $\alpha$  may introduce some false positive merging and further study is required for solving this problem. Although MetaCluster 3.0 does not perform very well in this dataset, it can still out-perform AbundanceBin which fails to estimate the correct number of clusters (4 clusters) in all cases.

### 3.5 Comparison on Running Time and Memory Consumption

Table 3 shows the running time and memory consumption of MetaCluster 3.0, MetaCluster 2.0 and AbundanceBin for different dataset sizes. The running times of the three algorithms increase with the input sizes. The running time of MetaCluster 3.0 and MetaCluster 2.0 are similar and much shorter than the running time of AbundanceBin, as AbundanceBin is required to construct a model for the distribution of reads and to repeat clustering the reads to estimate the number of clusters. The memory consumption of the three algorithms also increases with the input size but MetaCluster 3.0 consumes the least amount of memory.

## 4 CONCLUSIONS

In this paper, we propose a 2-phase (top-down separation and bottom-up merging) unsupervised binning algorithm to bin metagenomic fragments with mixed species abundance ratios. Based on the differences in the distribution of a distance measure between fragments of the same species and fragments from different species, our approach can guarantee the quality of our resulting clusters. The performance of our approach, MetaCluster 3.0, is shown to be better than all existing unsupervised algorithms. However, binning metagenomic fragments remains a challenging problem. All existing algorithms (including MetaCluster 3.0) can only handle datasets with not too many species and the accuracy decrease sharply when the number of species over ten. In the practical situations, a sample may contain genomes of thousands of kinds of species for which all existing binning tools fail.

There is another limitation of MetaCluster 3.0, which only works on fragments with length at least 500 nt. As the current high-throughput sequencing technology can only produce reads with lengths from 50 nt to 150 nt only, MetaCluster 3.0 relies on assembly tools for producing high-quality contigs with longer lengths. However, some binning algorithms, e.g. AbundanceBin, can work directly on short reads. Further research is required to come up with an effective tool for binning short reads directly with

mixed species abundance ratio or assembling reads in metagenomic data accurately.

## REFERENCES

- Altschul, S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25, 3389-3402.
- Amann, R.L. et al. (1990) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Environ Microbiol*, 56, 1919-1925.
- Case, R.J. et al. (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol* 73, 278-288.
- Chan, C.K. et al. (2008) Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics*, 9, 215.
- Chatterji, S. et al. (2008) A DNA composition-based algorithm for binning environmental shotgun reads. In *RECOMB, Proceedings*, 17-28.
- Chor, B. et al. (2009) Genomic DNA  $k$ -mer spectra: models and modalities. *Genome Biology*, 10, R108.
- Cole, J.R. et al. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res*, 33, D294-296.
- 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.
- Garcia Martin, H. et al. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol*, 24, 1263-1269.
- Huson, D.H. et al. (2007) MEGAN analysis of metagenomic data. *Genome Res*, 17, 377-386.
- Jain, A.K. and Dubes R.C. (1981) *Algorithms for Clustering Data*. Prentice-Hall.
- Jones, B.V. et al. (2008) Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc Natl Acad Sci USA*, 105, 13580-13585.
- Karlin, S. and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, 11, 283-290.
- Karlin, S. and Ladunga, I. (1994) Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci USA*, 91, 12832-12836.
- Karlin, S., Mrazek, J. and Campbell, A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *Jour Bacteriol*, 179, 3899-3913.
- Kendall, M.G. (1938) A new measure of rank correlation. *Biometrika*, 30, 81-93.
- Khachatryan, Z.A. et al. (2008) Predominant role of host genetics in controlling the composition of gut microbiota. *PLoS One*, 3, e3064.
- Kislyuk, A. et al. (2009) Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics*, 10, 316.
- Mavromatis, K. et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, 4, 495-500.
- Qin, J. et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464, 59-65.
- Teeling, H. et al. (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.
- Tringe, S.G. et al. (2005) Comparative metagenomics of microbial communities. *Science*, 308, 554-557.
- Tyson, G.W. et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428, 37-43.
- Venter, J.C. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304, 66-74.
- Wu, Y.W. and Ye, Y. (2010) A novel abundance-based algorithm for binning metagenomic sequences using  $l$ -tuples. In *RECOMB*, 535-549.
- Yang, B. et al. (2009) Unsupervised binning of environmental genomic fragments based on an error robust selection of  $l$ -mers. In *DTMBIO '09*, 3-10.
- Yang, B. et al. (2010) MetaCluster: Unsupervised Binning of Environmental Genomic Fragments and Taxonomic Annotation. *ACM-BCB*, 170-179.
- Zhou, F., Olman, V. and Xu, Y. (2008) Barcodes for genomes and applications. *BMC Bioinformatics*, 9, 546.