A Probabilistic Approach to Accurate Abundance-Based Binning of Metagenomic Reads

Olga Tanaseichuk[1], James Borneman[2], and Tao Jiang[1]

[1] Department of Computer Science and Engineering, University of California, Riverside, CA
[2] Department of Plant Pathology and Microbiology, University of California, Riverside, CA
{tanaseio,jiang}@cs.ucr.edu, borneman@ucr.edu

Abstract. An important problem in metagenomic analysis is to determine and quantify species (or genomes) in a metagenomic sample. The identification of phylogenetically related groups of sequence reads in a metagenomic dataset is often referred to as binning. Similarity-based binning methods rely on reference databases, and are unable to classify reads from unknown organisms. Composition-based methods exploit compositional patterns that are preserved in sufficiently long fragments, but are not suitable for binning very short next-generation sequencing (NGS) reads. Recently, several new metagenomic binning algorithms that can deal with NGS reads and do not rely on reference databases have been developed. However, all of them have difficulty with handling samples containing low-abundance species. We propose a new method to accurately estimate the abundance levels of species based on a novel probabilistic model for counting l-mer frequencies in a metagenomic dataset that takes into account frequencies of erroneous l-mers and repeated l-mers. An expectation maximization (EM) algorithm is used to learn the parameters of the model. Our algorithm automatically determines the number of abundance groups in a dataset and bins the reads into these groups. We show that our method outperforms the most recent abundance-based binning method, AbundanceBin, on both simulated and real datasets. We also show that the improved abundance-based binning method can be incorporated into a recent tool TOSS, which separates genomes with similar abundance levels and employs AbundanceBin as a preprocessing step to handle different abundance levels, to enhance its performance. We test the improved TOSS on simulated datasets and show that it significantly outperforms TOSS on datasets containing low-abundance genomes. Finally, we compare this approach against very recent metagenomic binning tools MetaCluster 4.0 and MetaCluster 5.0 on simulated data and demonstrate that it usually achieves a better sensitivity and breaks fewer genomes.

Keywords: metagenomics, next-generation sequencing, expectation maximization, abundance-based binning.

B. Raphael and J. Tang (Eds.): WABI 2012, LNBI 7534, pp. 404–416, 2012.

1 Introduction

Metagenomics studies the genomic content of an entire microbial community by simultaneously sequencing all genomes in an environmental sample. This approach allows us to study previously uncultured microorganisms that constitute the vast majority of organisms in most environmental and clinical samples [1]. Metagenomics has already led to a better understanding of microbial communities in various environments, *e.g.* acid-mine drainage ponds [2], human gut [3], soil [4], and marine worms [5]. The recent advent of *next-generation sequencing (NGS)* technologies [6,7] has drastically improved sequencing time and cost, leading to an exponential increase in environmental sequencing data which makes it possible to study microbial communities at a much higher resolution due to increased sequencing depth [8]. NGS-based approaches have recently been applied to sequence several metagenomes from cow rumen [9], saliva microbiome [10], permafrost [11], etc.

In metagenomics, a sample contains sequence reads from various organisms. Therefore, an important problem in a metagenomic analysis is to determine and quantify the species (or genomes) in a sample. The identification of phylogenetically related groups of reads in a metagenomic dataset is usually referred to as *binning*. A handful of binning algorithms have been developed for metagenomic datasets. Similarity-based methods explore the taxonomic composition of metagenomic sequences by performing similarity search against databases of known genomes, genes and proteins [12,13,14,15]. These methods have high accuracy and are suitable for very short NGS reads. However, they rely on the availability of reference databases, while a lot of organisms in a sample may not be remotely related to any known species. As a consequence, a large fraction of read data may remain unclassified.

Another group of binning methods is based on compositional properties of the reads. These methods rely on the property that compositional features, such as oligonucleotide frequencies and CG content, are preserved across sufficiently long fragments of a genome. Supervised composition-based algorithms exploit compositional properties of the reads for taxonomic classification against models trained on known sequences [16,17,18]. Unsupervised methods perform clustering of the reads to detect groups of reads from related organisms [19,20,21,22]. Composition-based methods can accurately bin long fragments. However, due to local variation of DNA composition across a genome, the performance of these methods degrades with the decrease of the read length, making them unsuitable for NGS datasets.

Several recent unsupervised metagenomic binning algorithms have been developed to handle short NGS reads. In particular, MetaCluster 4.0 [23] exploits compositional properties of groups of reads rather than individual reads. Although it handles high-abundance species well, it does not perform well on datasets with low-abundance species. MetaCluster 5.0 [24] is a very recent extension of MetaCluster 4.0 to deal with low-abundant species. Another unsupervised binning algorithm that handles NGS reads is AbundanceBin [25]. It is designed to separate reads from genomes with different abundance levels. To predict the

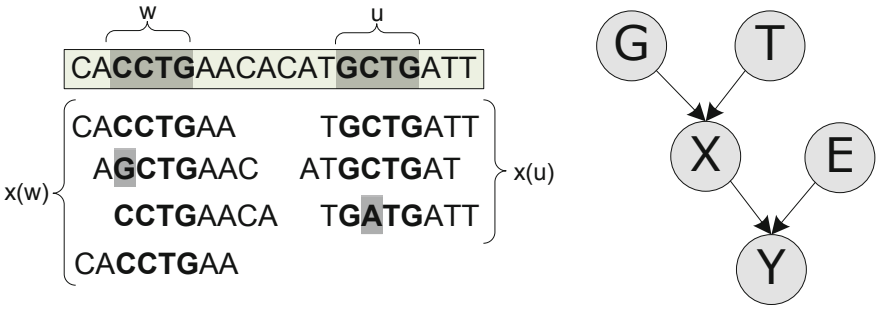
abundance levels, the frequencies of l-mers are modeled as a mixture of Poisson distributions. In this model, repeated l-mers and l-mers with errors are ignored, which may often lead to an inaccurate estimation of the parameters and result in a low binning accuracy. TOSS [26] is designed to separate reads from genomes with similar abundance levels. In the first phase, TOSS creates clusters of l-mer so that all l-mer in each cluster are likely to originate from the same genome. In the second phase, clusters from the same genome are merged. When genomes have different abundance levels, TOSS uses AbundanceBin as a preprocessing step. Clearly, the performance of TOSS is significantly affected by the performance of AbundaceBin. Specifically, the inability of AbundaceBin to accurately infer low-coverage genomes may result in bins with low sensitivity. When these bins are provided to TOSS as an input, the performance of TOSS would suffer. To address this problem, we introduce a method to accurately determine the abundance levels of genomes in a metagenomic dataset.

In this paper, we propose a novel probabilistic model for counting l-mer frequencies in the metagenomic dataset that takes into account the frequencies of erroneous l-mers as well as repeats. An *expectation maximization (EM)* algorithm is used to learn the parameters of the model. The algorithm automatically determines the number of abundance groups in the dataset and bins the reads into these groups. We show that the method outperforms AbundanceBin on siumlated and real datasets. We also show that the method can be incorporated into TOSS to improve its performance in the presence of genomes with different abundance levels. In fact, our experiments on simulated datasets demonstrate that this method significantly improves the performance of TOSS. Finally, we compare the improved TOSS against recent metagenomic binning tools MetaCluster 4.0 and MetaCluster 5.0 on simulated NGS datasets, and show that it has a comparable performance overall but often achieves a better sensitivity and breaks fewer genomes.

The paper is organized as follows. In section 2, we describe the probabilistic model for counting l-mer frequencies in a metagenomic dataset and the algorithm for learning the parameters of the model and automatically detecting the number of abundance groups. Section 3 presents the experimental evaluation of our method and comparison to other recent binning methods. Section 4 concludes the paper.

2 Methods

In this section, we introduce a novel probabilistic model that can be used for computing the most probable abundance levels of the genomes in a metagenomic dataset and estimating the proportions of the reads corresponding to each abundance group. The problem of binning the reads is then reduced to the problem of determining the parameters of the model and classifying the reads according to the frequencies of l-mers comprising the reads.



(a) The coverage of l-mer w = CCTG is $x(w) = 4$. However, due to an error in one of the reads that cover w, w appears in the reads only 3 times, *i.e.* $y(w) = 3$. For the l-mer u = GCTG, $x(u) = 3$. Observe that even though there is an error in one of the reads that cover u, this l-mer also occurs in a read that covers w due to an error, and thus $y(u) = 3$.

(b) Count Y of an l-mer depends on the coverage X and the number of errors E within the l-mer. In turn, the coverage depends on the abundance level G of the genome and the number of occurances T of the l-mer in the genome.

Fig. 1. Left: Coverage of l-mers and occurrences of l-mers in the reads. Right: The proposed graphical model.

2.1 Definitions and Notations

Assume that N reads are drawn randomly from a genome of length L_g. Let L be the length of a read. According to the Lander-Waterman model [27], the left ends of the reads can be modeled by a Poisson process. Under this model, the number of reads that cover each substring of length l of the genome follows a Poisson distribution with the parameter $\lambda = N(L - l + 1)/(L_g - L + 1)$. From now on, we will refer to λ as *the abundance level of the genome*.

Even though most of the l-mers in a bacterial genome occur only once within the genome [26], some l-mers may occur at multiple locations within the genome. Assume that w is an l-mer with n copies in the genome. Due to additivity of the Poisson distribution, the number of reads that cover w, denoted by $x(w)$, has a Poisson distribution with the parameter $n\lambda$. However, due to sequencing errors, the actual count of the l-mer w in the reads, denoted by $y(w)$, may differ from $x(w)$ (see Figure 1a). Let $x^i(w)$ be the number of reads that cover the l-mer w with i errors in w. Clearly, $x(w) = \sum_i x^i(w)$ and $y(w) = x^0(w) + e_w$, where e_w is the number of times that w occurs in the reads due to errors in other l-mers.

Now, let us consider a metagenomic dataset. Assume that N reads are sequenced from S different genomes. The abundance value of genome g_j is

$\lambda_j = N_j(L-l+1)/(L_{g_j}-L+1)$, where N_j is the number of reads corresponding to this genome and L_{g_j} is the length of the genome g_j. Let us enumerate all the substrings of length l in all the reads. Clearly, there are $M = N(L-l+1)$ such substrings. Let us consider the i^{th} substring v_i, $i \in [1, M]$. This substring belongs to the read $r_i \in [1, N]$ which was sequenced from the genome $g_i \in [1, S]$. Let w_i be the original l-mer in the genome g_i corresponding to v_i. Let us assume that w_i has t_i copies in genome g_i. Let e_i be the number of sequencing errors (substitutions) within v_i. Note that e_i equals the Hamming distance between w_i and v_i. Also, let x_i be the number of reads that cover all the copies of w_i in the genome and y_i the number of times that l-mer v_i occurs in the reads.

Next, we model the relationship between the abundance values of genomes, the coverage of l-mers, the number of errors in l-mers, and the counts of l-mers in the reads.

2.2 A Probabilistic Model for l-mer Frequencies

We define random variables G_i, X_i, Y_i, T_i, and E_i that are associated with the values g_i, x_i, y_i, t_i and e_i, respectively. The variables Y_i are observed by counting the number of occurrences of l-mers in the reads. The other variables cannot be observed directly, so they are hidden. Our goal is to determine the most likely assignment of the l-mers to the genomes. Figure 1b illustrates a graphical representation of the model.

Let π_j be a parameter that represents the proportion of the reads that come from the j^{th} genome. Let α_j^n be the fraction of l-mers that occur n times in the j^{th} genome. Let $\alpha_j = (\alpha_j^1, ..., \alpha_j^{n_{max}})$, where n_{max} is the maximum possible number of copies of an l-mer in a genome. For the convenience of notation, we define parameter vectors $\theta_j = (\lambda_j, \pi_j, \alpha_j)$ for all $j \in [1, S]$, and $\theta = (\theta_1, ..., \theta_S)$.

Assuming that the coverage of an l-mer with t copies in a genome g follows a Poisson distribution, the probability that the random variable X_i, associated with the coverage of l-mers in the genome g, takes a particular value c is

$$P(X_i = c | G_i = g, T_i = t, \theta) = \frac{c\,\mathrm{Pois}(t\lambda_g, c)}{\sum_j j\,\mathrm{Pois}(t\lambda_g, j)} = \mathrm{Pois}(t\lambda_g, c-1),$$

where $\mathrm{Pois}(\lambda, k)$ is the probability of a Poisson random variable taking the value k.

The variable Y_i associated with the count of the l-mer v_i in the reads conditionally depends on variables X_i and E_i. If $E_i = e, e > 0$, it means that the corresponding l-mer v_i contains e errors. To model the distribution of counts of l-mers that have e errors, we can borrow the idea from the Balls and Bins problem (`http://www.mathpages.com/home/kmath199.htm`).

Assume that n balls are randomly thrown into m bins. It is known that the expected fraction of bins that get exactly k balls can be approximated by a Poisson distribution with the parameter n/m. Based on this, the probability that an erroneous l-mer has frequency k in the reads is

$$P(Y_i = k|X_i = c, E_i = e, e > 0, \theta) = \frac{kc\,\mathrm{Pois}(c/n_l(e), k)}{\sum_j j\,\mathrm{Pois}(c/n_l(e), j)} = \mathrm{Pois}(c/n_l(e), k-1),$$

where $n_l(e)$ is the number of different possibilities for e errors to occur within an l-mer.

The distribution of the counts of l-mers without errors can be modeled by the binomial distribution. The probability that an error-free l-mer has count k in the reads is

$$P(Y_i = k|X_i = c, E_i = 0) = \frac{k\,\mathrm{Bin}(k, c, p_0)}{\sum_j j\,\mathrm{Bin}(j, c, p_0)} = \frac{k\,\mathrm{Bin}(k, c, p_0)}{cp_0}$$

where $\mathrm{Bin}(j, c, p)$ is the probability that a variable following the binomial distribution takes the value j, and p_0 is the probability that an l-mer does not contain erros.

The above probabilities allow us to compute the probability of a given data point y_i given the values of unobserved variables and the parameter vector θ

$$\begin{aligned}
&P(Y_i = y_i|G_i = g, X_i = c, E_i = e, T_i = t, \theta) \\
&= P(Y_i = y_i|X_i = c, E_i = e)P(X_i = c|G_i = g, T_i = t, \theta)P(G_i = g, T_i = t, \theta)P(E_i = e) \\
&= \pi_g \alpha_g^t P(Y_i = y_i|X_i = c, E_i = e)P(X_i = c|G_i = g, T_i = t, \theta)P(E_i = e) \qquad (1)
\end{aligned}$$

2.3 Parameter Estimation

Now, let us consider the log-likelihood of the observed data Y given the parameter vector θ

$$L(Y|\theta) = \sum_i \log P(Y_i = y_i|\theta).$$

Our goal is to find the *maximum likelihood estimate (MLE)* of the parameter θ,

$$\hat{\theta} = \arg\max_\theta L(Y|\theta).$$

To find $\hat{\theta}$, we use the EM algorithm. The E-step requires the computation of the expected value of the log-likelihood function, with respect to the conditional distribution of unobservable variables given the data and current parameter estimates $\theta^{(t)}$:

$$\begin{aligned}
Q(\theta|\theta^{(n)}) = \sum_i \sum_{g,c,e,t} &P(G_i = g, X_i = c, T_i = t, E_i = e|Y_i = y_i, \theta^{(t)}) \\
&\cdot P(Y_i = y_i, G_i = g, X_i = c, T_i = t, E_i = e|\theta)
\end{aligned}$$

Here, the posterior probabilities $p_{G,X,E,T|Y,\theta}(g, c, e, t, k, \theta) = P(G_i = g, X_i = c, E_i = e, T_i = t|Y_i = k, \theta)$ of the unobserved data given current parameter estimates $\theta^{(t)}$ can be computed by applying Bayes' rule to Equation 1.

In the M-step, we find the parameter $\theta^{(t+1)}$ that maximizes $Q(\theta|\theta^{(n)})$ with respect to θ

$$\theta^{(n+1)} = \arg\max_{\theta} Q(\theta|\theta^{(n)}). \tag{2}$$

The updated parameters are thus

$$\lambda_g^{(n+1)} = \frac{\sum_{c,e,t,k} p_{G,X,E,T|Y,\theta}(g,c,e,t,k,\theta)c}{\sum_{c,e,t,k} p_{G,X,E,T|Y,\theta}(g,c,e,t,k,\theta)t}, \; \alpha_g^{t\,(n+1)} = \frac{\sum_{c,e,k} p_{G,X,E,T|Y,\theta}(g,c,e,t,k,\theta)}{\sum_{c,e,j,k} p_{G,X,E,T|Y,\theta}(g,c,e,j,k,\theta)}$$

$$\pi_g{}^{(n+1)} = \frac{\sum_{c,e,k,j} p_{G,X,E,T|Y,\theta}(g,c,e,j,k,\theta)}{\sum_{i,c,e,j,k} p_{G,X,E,T|Y,\theta}(i,c,e,j,k,\theta)}$$

Once we estimate the parameters of the probabilistic model, we can assign l-mers to bins (or genomes) based on the counts of the l-mers in the reads. We assign an l-mer v_i that occurs y_i times in the reads to a bin g with probability $P(G_i = g|Y_i = y_i, \hat{\theta})$. Then, each read is assigned to a bin according to the frequencies of its l-mers in the dataset

$$P(r \in g_j) = \prod_{y_i \in r} P(G_i = g|Y_i = y_i, \hat{\theta}) / \sum_g \prod_{y_i \in r} P(G_i = g|Y_i = y_i, \hat{\theta}).$$

2.4 Detecting the Number of Bins

The EM algorithm described above assumes that the number of bins (genomes) S and the maximum multiplicity of the repeats in the genome (the values that variables T_i may take) are provided. Selecting the best number of clusters is a challenging problem. Here, we propose an iterative algorithm to find the best value for S. We start with one bin and iteratively increase the number of bins until one of the following conditions is reached: (i) one or several bins are split into overlapping bins, making it impossible to assign the reads to the overlapping bins correctly and (ii) one or several bins are too small to represent a whole genome. In order to find the maximum multiplicity of the repeats, denoted by R, we repeat the above procedure for different values of R. For each pair of specific values $S = s$ and $R = r$, we record the distance between the observed and the expected frequencies of l-mers, $V(s,r) = \sum_i |M \cdot P(Y = i|\hat{\theta}_{r,s}) - \sum_{j=1..M} \mathbb{1}_{\{i\}}(y_j)|$. Here $M \cdot P(Y = i|\hat{\theta}_{r,s})$ is the expected number of l-mers with counts i, and $\sum_{j=1..M} \mathbb{1}_{\{i\}}(y_j)$ is the observed number of l-mers with counts i in the reads. Finally, we set S and R to the values s and r for which $V(s,r)$ reaches the minimum. See Algorithm 1 below for the details.

3 Experimental Results

We demonstrate the performance of our abundance-based binning algorithm on simulated and real datasets and compare the results with AbundanceBin. We also

Algorithm 1: Deciding the optimal number of bins S and maximum multiplicity of the repeats R. Given observed l-mer frequencies, the algorithm attempts to find the best values for S and R.

```
begin
    V ← ∞
    R, S ← 1, 1
    for r = 1, ..., R_max do
        s ← 1
        θ̂ ← EM(s, r)
        if StopCondition(θ̂) then
            break
        else
            if V(s, r) < V then
                V ← V(s, r)
                R, S ← r, s
        s ← s + 1
    return R, S
end
```

show that the algorithm can be incorporated into our recent genome separation tool TOSS to enhance its performance. We test the improved TOSS on simulated NGS datasets and compare the results with those of TOSS that uses (or does not use) AbundanceBin as a preprocessor. Finally, we compare the performance of the improved TOSS with two very recent binning tools MetaCluster 4.0 and MetaCluster 5.0 on simulated NGS data.

3.1 Performance on a Simulated Data

Due to our limited knowledge of the nature of microbial communities, simulated metagenomic datasets are widely used for testing the performance of existing metagenomic tools. We simulated several metagenomic datasets based on complete genomes from the NCBI database using software MetaSim [28]. Each simulated dataset contains paired-end reads of length 80 bps. The sequencing error model was set according to the Illumina error profile with 1% average sequencing error rate.

We compare the performance of our algorithm against AbundanceBin. In this test, we are mainly concerned with the ability of both algorithms to separate reads from genomes with different abundance levels. In order to measure the performance of the algorithms, we use the evaluation criteria defined in [26]. We assign a genome to a bin (or cluster) if more than half of the reads from the genome are assigned to this bin. If there is no bin that contains the majority of the reads from a genome, we report the genome as broken. We allow several genomes to be assigned to one bin, and say that the genomes are not separated if the reads of the genomes ended up in the same cluster. We compute the *separability rate* as the percentage of separated pairs of genomes in the dataset.

Table 1. Comparison with AbundanceBin on simulated datasets. The bold numbers indicate improved sensitivity and precision. The numbers in parentheses are normalized sensitivity and precision.

ID	# genomes	Cove-rage	Length Mbp	Ours			AbundanceBin		
				Sens.	Prec.	Sep.	Sens.	Prec.	Sep.
S1	2	5 10	2.0 1.9	0.80 **(0.84)**	**0.84 (0.84)**	1	0.80 (0.75)	0.77 (0.76)	1
S2	3	5 5 11	2.7 2.6 3.0	**0.89 (0.89)**	**0.89 (0.89)**	1	0.86 (0.85)	0.86 (0.85)	1
S3	2	5 9	0.6 0.6	**0.79 (0.81)**	**0.78 (0.80)**	1	0.74 (0.69)	0.74 (0.69)	1
S4	2	4 8	4.4 5.2	**0.73 (0.82)**	**0.81 (0.81)**	**1**	-	-	0
S5	3	3 3 8	5.7 4.4 6.0	0.87 **(0.93)**	**0.88 (0.93)**	1	**0.91** (0.89)	0.80 (0.89)	1
S6	3	3 8 15	4.6 4.1 4.7	0.75 (0.83)	0.83 (0.82)	**1**	**0.81 (0.84)**	**0.88 (0.84)**	0.66
S7	6	2,2 2,6 6,6	1.5,1.8 2.0,1.7 1.8,2.0	**0.86 (0.75)**	**0.85 (0.84)**	**1**	-	-	0

In addition to standard sensitivity and precision, we also measure *normalized sensitivity* and *precision*. The formal definitions of these concepts can be found in [26].

The detailed datasets and performance of the two algorithms are summarized in Table 1. On most of the datasets, the sensitivity and precision of our method were better than those of AbundanceBin by 4-10%. In tests S4 and S7, AbundanceBin failed to separate the two genomes totally. In test S6, AbundanceBin could identify only 2 bins, while combining the reads from two genomes into one bin. Our method was more ambitious and separated all three genomes at the cost of lowered precision and sensitivity. However, when we set the number of bins to two for the dataset in test S6, our algorithm was able to achieve a high sensitivity and precision above 95%, compared to 81% and 88% for AbundanceBin.

3.2 Performance on a Real Dataset

We test the performance of our method on a dataset obtained from the acid mine drainage [2]. This dataset has been well studied and is known to contain five dominant genomes. The two most abundant species belong to *Leptospirillum* group II and *Ferroplasma* group II. The three species with a lower abundance levels belong to *Leptospirillum* group III, *Ferroplasma* group I and *Sulfobacill*. The dataset consists of approximately 120K Sanger reads. Only 56% percent of the reads can be mapped to the reference sequences of the five dominant genomes.

Table 2. Performance of the improved TOSS and comparison with the previous TOSS, MetaCluster 4.0 and MetaCluster 5.0. The bold numbers indicate the best performance among all five methods.

ID	# genomes	Coverage	Ours+TOSS				ABin+TOSS				TOSS				MC 4.0				MC 5.0			
			Sens.	Prec.	Sep.	Broken	Sens.	Prec.	Sep.	Broken	Sens.	Prec.	Sep.	Broken	Sens.	Prec.	Sep.	Broken	Sens.	Prec.	Sep.	Broken
T1	4	4,4, 10,10	**1.0**	**0.84**	**1.0**	**0**	-	-	0	0	0.63	0.55	**1.0**	**0**	-	-	-	1	0.75	0.69	**1.0**	**0**
T2	3	4,10,10	**1.0**	0.96	**1.0**	**0**	0.62	0.72	**1.0**	**0**	0.73	0.84	**1.0**	**0**	0.91	**1.0**	**1.0**	**0**	0.79	**1.0**	**1.0**	**0**
T3	3	4,12,12	**1.0**	**1.0**	**1.0**	**0**	**1.0**	0.99	**1.0**	**0**	0.84	0.90	**1.0**	1	0.97	0.96	**1.0**	**0**	0.82	**1.0**	**1.0**	**0**
T4	4	7,7,13,13	0.86	0.82	0.83	**0**	0.76	0.76	0.67	**0**	-	-	0	2	**0.89**	1.0	**1.0**	2	0.84	**1.0**	**1.0**	**0**
T5	10	1,1,1,2, 2,2,1.5, 1.5,10,10	**1.0**	0.92	0.83	**0**	-	-	0	0	**1.0**	0.64	0	2	0.78	**0.97**	**1.0**	**0**	-	-	-	4
T6	10	1.5,1.5,1.5, 1.5,1.5,1.5, 9,9,9,9	0.91	0.87	**1.0**	**0**	-	-	0	2	**0.99**	0.81	**1.0**	**0**	0.80	0.96	**0**	0	0.74	**1.0**	1.0	2
T7	18	2,2,2,2, 3,3,3,3, 3,3,3,4, 4,4, 11, 12,12,12	0.87	0.75	0.73	**0**	-	-	0	0	-	-	-	3	**0.9**	**0.9**	**1**	**0**	0.88	**0.9**	**1**	**0**

We apply both our algorithm and AbundanceBin to the unfiltered dataset. Then we BLAST the reads of each bin against reference sequences of the five organisms. We measure the ability of the algorithms to separate reads from the two main abundance groups. Although both algorithms could correctly identify the two bins, our algorithm slightly outperforms AbundanceBin in terms of precision and sensitivity. Our method achieves 82% sensitivity and 81% precision, while the corresponding values are 78% and 79% for AbundanceBin. Note that due to the overlap of the bins, it would be very difficult to separate the reads with much better sensitivity and precision based on l-mer frequencies only.

3.3 Performance of the Improved TOSS

TOSS is designed to handle genomes with similar abundance levels and it requires a preprocessing step to separate the reads from the genomes with different abundance levels. We incorporate our abundance-based binning algorithm into TOSS and test the performance of the improved TOSS on simulated NGS datasets. We compare the results with the previous version of TOSS that employs AbundanceBin as a preprocessor and with TOSS without any preprocessing steps. Also, we make a comparison with the most recent metagenomic NGS binning tools MetaCluster 4.0 and MetaCluster 5.0. Again, to measure the performance of the tools, we use the evaluation criteria defined in [26]. The results of the comparison are summarized in Table 2. Note that here we only measure the ability of the algorithms to separate high-abundance genomes (with abundance levels ≥ 7, as done in [24]). The improved TOSS obviously outperforms both the version of TOSS that relies on AbundanceBin and the version of TOSS that does not use any preprocessor (the former has low separability rate while the latter yields a high number of broken genomes). Compared to the MetaCluster tools, our algorithm often achieves the highest sensitivity and breaks fewer genomes.

4 Conclusion

Metagenomics approach has opened a door into the previously hidden world of microorganisms. However, analysis of metagenomic data remains a difficult problem far from being solved. Binning is an important step of metagenomic analysis. In this paper, we introduced a novel probabilistic model for counting l-mer frequencies in a metagenomic dataset. The model allows us to identify the most probable abundance levels of the genomes in a metagenomic sample accurately and estimate the proportions of reads from corresponding genomes. We have shown that our model can serve as a useful preprocessing tool for further metagenomic analysis.

Acknowledgments. We are grateful to the anonymous referees for their many constructive comments. The research was supported in part by NIH grant AI078885.

References

1. Amann, R.I., Ludwig, W., Schleifer, K.H.: Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews 59(1), 143–169 (1995)
2. Tyson, G.W., Chapman, J., Hugenholtz, P., et al.: Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978), 37–43 (2004)
3. Gill, S.R., Pop, M., DeBoy, R.T., et al.: Metagenomic Analysis of the Human Distal Gut Microbiome. Science 312(5778), 1355–1359 (2006)
4. Tringe, S.G., von Mering, C., Kobayashi, A., et al.: Comparative Metagenomics of Microbial Communities. Science 308(5721), 554–557 (2005)
5. Woyke, T., Teeling, H., Ivanova, N.N., et al.: Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443(7114), 950–955 (2006)
6. Margulies, M., Egholm, M., Altman, W.E., et al.: Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057), 376–380 (2005)
7. Bentley, D.R.: Whole-genome re-sequencing. Current opinion in genetics & development 16(6), 545–552 (2006)
8. Singh, A.H., Doerks, T., Letunic, I., et al.: Discovering Functional Novelty in Metagenomes: Examples from Light-Mediated Processes. J. Bacteriol. 191(1), 32–41 (2009)
9. Hess, M., Sczyrba, A., Egan, R., et al.: Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331(6016), 463–467 (2011)
10. Yang, F., Zeng, X., Ning, K., et al.: Saliva microbiomes distinguish caries-active from healthy human populations. The ISME Journal 6(1), 1–10 (2011)
11. Mackelprang, R., Waldrop, M.P., DeAngelis, K.M., et al.: Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature 480(7377), 368–371 (2011)
12. Huson, D.H., Auch, A.F., Qi, J., et al.: MEGAN analysis of metagenomic data. Genome research 17(3), 377–386 (2007)
13. Krause, L., Diaz, N.N., Goesmann, A., et al.: Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Research 36(7), 2230–2239 (2008)
14. Ghosh, T., Monzoorul Haque, M., Mande, S.: DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. BMC Bioinformatics 11(suppl. 7), S14+ (2010)
15. Monzoorul Haque, M., Ghosh, T.S.S., Komanduri, D., Mande, S.S.: SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics (Oxford, England) 25(14), 1722–1730 (2009)
16. Diaz, N., Krause, L., Goesmann, A., et al.: TACOA - Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics 10(1), 56+ (2009)
17. McHardy, A.C., Martin, H.G., Tsirigos, A., et al.: Accurate phylogenetic classification of variable-length DNA fragments. Nature Methods 4(1), 63–72 (2006)
18. Brady, A., Salzberg, S.L.: Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat. Meth. 6(9), 673–676 (2009)
19. Chatterji, S., Yamazaki, I., Bai, Z., et al.: CompostBin: A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads. In: Vingron, M., Wong, L. (eds.) RECOMB 2008. LNCS (LNBI), vol. 4955, pp. 17–28. Springer, Heidelberg (2008)

20. Teeling, H., Waldmann, J., Lombardot, T., et al.: TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5(1), 163+ (2004)
21. Prabhakara, S., Acharya, R.: A two-way multi-dimensional mixture model for clustering metagenomic sequences. In: Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine, BCB 2011, pp. 191–200. ACM (2011)
22. Yang, B., Peng, Y., Leung, H., et al.: Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. BMC Bioinformatics 11(Suppl 2), S5+ (2010)
23. Wang, Y., Leung, H.C., Yiu, S.M., Chin, F.Y.: MetaCluster 4.0: A Novel Binning Algorithm for NGS Reads and Huge Number of Species. Journal of Computational Biology: a Journal of Computational Molecular Cell Biology 19(2), 241–249 (2012)
24. Wang, Y., Leung, H., Yiu, S., Chin, F.: Metacluster 5.0: A two-round binning approach for metagenomic data for low-abundance species in a noisy sample. In: Proceedings of the ECCB (to appear, 2012)
25. Wu, Y.-W., Ye, Y.: A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using *l*-Tuples. In: Berger, B. (ed.) RECOMB 2010. LNCS, vol. 6044, pp. 535–549. Springer, Heidelberg (2010)
26. Tanaseichuk, O., Borneman, J., Jiang, T.: Separating Metagenomic Short Reads into Genomes via Clustering. In: Przytycka, T.M., Sagot, M.-F. (eds.) WABI 2011. LNCS, vol. 6833, pp. 298–313. Springer, Heidelberg (2011)
27. Lander, E.S., Waterman, M.S.: Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2(3), 231–239 (1988)
28. Richter, D.C., Ott, F., Auch, A.F., et al.: MetaSim: a Sequencing Simulator for Genomics and Metagenomics. PLoS ONE 3(10), e3373+ (2008)