Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 153647, 11 pages
doi:10.1155/2012/153647

Research Article

Unsupervised Two-Way Clustering of Metagenomic Sequences

Shruthi Prabhakara and Raj Acharya

Department of Computer Science and Engineering, University Park, PA 16802, Pennsylvania State University, USA

Correspondence should be addressed to Shruthi Prabhakara, sap263@psu.edu

Received 15 December 2011; Accepted 26 January 2012

Academic Editor: Wei Wang

A major challenge facing metagenomics is the development of tools for the characterization of functional and taxonomic content of vast amounts of short metagenome reads. The efficacy of clustering methods depends on the number of reads in the dataset, the read length and relative abundances of source genomes in the microbial community. In this paper, we formulate an unsupervised naive Bayes multispecies, multidimensional mixture model for reads from a metagenome. We use the proposed model to cluster metagenomic reads by their species of origin and to characterize the abundance of each species. We model the distribution of word counts along a genome as a Gaussian for shorter, frequent words and as a Poisson for longer words that are rare. We employ either a mixture of Gaussians or mixture of Poissons to model reads within each bin. Further, we handle the high-dimensionality and sparsity associated with the data, by grouping the set of words comprising the reads, resulting in a two-way mixture model. Finally, we demonstrate the accuracy and applicability of this method on simulated and real metagenomes. Our method can accurately cluster reads as short as 100 bps and is robust to varying abundances, divergences and read lengths.

1. Introduction

Metagenomics is defined as the study of genomic content of microbial communities in their natural environments, bypassing the need for isolation and laboratory cultivation of individual species [1]. Its importance arises from the fact that over 99% of the species yet to be discovered are resistant to cultivation [2]. This limitation imposed by cultivation of isolated clones has severely skewed our view of microbial diversity. Metagenomics promises to enable scientists to study the full diversity of the microbial world, their functions and evolution, in their natural environments.

Next Generation Sequencing (NGS) technologies generate data more efficiently, economically, and with a greater depth than ever before. NGS has opened up an array of possibilities for many applications including whole-genome sequencing, epigenetics, and metagenomics. Of these, the characterization of diversity of heterogeneous microbial environments, metagenomes, has recently gained significant interest. Although a host of methods for whole-genome assembly have been developed, reconstruction of individual clones from metagenomes still remains a challenge. As compared to existing technologies, reads produced by NGS are typically shorter and more error-prone. The growth in the size of the datasets is fast outpacing the computational power needed to analyze it. Thus, many computational challenges arise while analyzing deep sequence data from heterogeneous populations [3]. The computational method we present here aims to quantify the microbial diversity within a metagenome based on a set of deep sequencing reads.

In single genome sequencing, we can be certain that all extracted DNA fragments belong to the same genome. This makes sequence assembly and annotation tractable. However, in majority of metagenomic samples, it is not possible to isolate and culture individual clones. It is further complicated by the fact that the data comes from heterogeneous microbial communities, where the number of species as well as their relative abundance is unknown. Sequence data is usually partial and fragmentary, as environmental sequence sampling rarely produces all the sequences required for assembly. Many of these species do not have fully sequenced genomes available. Moreover, metagenomic datasets are beset with increased amounts of polymorphism and horizontal gene transfer. Sequences from closely related species will most

likely have homologous sequences shared between them, hindering their separation [4]. As a result, the reconstruction of a whole genome is generally not possible.

In the light of new data, we need to adapt the traditional approaches to analyze metagenomic sequences. An additional step in metagenomics that is not required in single genome assembly is binning the reads belonging to different species that is the need to associate the reads with its source organism. Clustering methods aim to identify the species present in the sample, classify the reads by their species of origin, and quantify the abundance of each of these species. Clustering provides deeper insight into the structure of the community. It can lead to faster and more robust assembly by reducing the search space [5].

Most of the existing classification methods are supervised and depend on the availability of reference data for training [4–9]. A metagenomic dataset may, however, contain reads from unexplored phyla which cannot be labeled into one of the existing classes. Most metagenomic analysis methods until now have been relatively inaccurate in classifying short reads. Poor performance on the short fragments is mostly due to the high dimensionality and sparsity associated with the data [10]. To overcome the limitation imposed by the length, the reads are often assembled into longer contigs and then clustered. However, there is the danger of assembling reads from different species together, thereby creating interspecies chimeras. The presence of highly conserved sequences further occludes cluster boundaries between species. Moreover, the abundances of different species can be potentially skewed such that the within-species variance overwhelms the between-species variance [10].

In this paper, we develop a method for clustering the short metagenome reads that addresses the challenges posed by the nature of metagenomic data. We formulate an unsupervised two-way multispecies, multidimensional mixture model to represent reads from a metagenome. We model the distribution of word counts along a genome as a Gaussian for shorter, frequent words and as a Poisson for longer words that are rare. We employ either a mixture of Gaussians or a mixture of Poissons to model reads within each bin. The proposed model is used to cluster metagenomic reads by their species of origin and to characterize the abundance of each species. Our method is unsupervised in that it does not require any training data. It is a composition-based method that seeks to distinguish between genomes based on their characteristic DNA compositional pattern. Our method handles the high-dimensionality and sparsity associated with the data by grouping the set of words comprising the reads, to regularize the parameters in the mixture model. This implies that, for every group, only one statistic of the words in this group is needed to cluster reads. We show that a high clustering accuracy can be obtained at a much lower dimension. We provide a framework that complements existing similarity-based methods. Later in the paper, we evaluate the applicability of the multidimensional mixture of distributions and its ability to estimate the parameters of genome abundance accurately, for simulated and real metagenomes. We compare the performance of our method with LikelyBin and Scimm, two other unsupervised composition-based method. Also, we demonstrate the robustness of our method to changes in the relative abundance of different species.

2. Related Work

The last decade has seen an explosion in the number of computational methods developed to analyze metagenomic data. Literature abounds in methods for classifying (as opposed to clustering) metagenome reads into taxon-specific bins [5, 7, 8]. Current approaches to metagenomics clustering can be classified into two main categories: similarity-based and composition-based methods.

The similarity-based approaches align the reads to close phylogenetic neighbors and hence depend on the availability of closely related genomes in existing databases [6, 11, 12]. MEGAN, a metagenome analysis software system [6], is a representative example of this kind. It uses sequence homology to assign reads to common ancestors based on best match as given by BLAST (Basic Local Alignment Search Tool) [13]. As most of the extant databases are highly biased in their representation of true diversity, such methods fail to find homologs for reads derived from novel species.

A second class of computational methods bin the reads based on DNA composition. These methods rely on the intrinsic features of the reads such as oligonucleotide distributions [7, 8, 14], codon usage preference [15], and GC composition [16] to differentiate between reads belonging to different species. These "genome signatures" are known to be fairly constant throughout the genome. A significant limitation of most composition-based methods developed so far is that they do not perform well on reads shorter than 500 bp. Composition-based clustering methods of metagenome reads complement those based on similarity.

Phylopythia [7] is a supervised composition-based classification method that trains a support vector machine to classify sequences of length greater than 1 kbp. Phymm uses interpolated Markov models to characterize variable length DNA sequences by their phylogenetic group [8]. Its accuracy of assignment drops drastically (to just 7.1% at genus level) for short reads and reads from unknown species. Nasser et al. [5] demonstrated that a k-means based fuzzy classifier, trained using a maximal order Markov chain, can separate fragments that are about 1 kbp long at the phylum level with a high accuracy. Rosen et al. trained a Naive Bayes classifier using publicly available microbial genomes [9]. CompostBin is a semisupervised algorithm for grouping fragments that uses a novel weighted PCA (Principal Component Analysis) and a normalized cut clustering algorithm to classify the sequences [10]. They have demonstrated an error rate bounded by 10%, when guided by information from phylogenetic markers, on datasets of low complexity. However, the accuracy of this method on reads less than 1 kbp has not been shown. Recently, Chan et al. developed a semisupervised seeded growing self-organizing map (S-GSOM) [4] to cluster metagenomic sequences. It extracts 8–13 kbp of flanking sequences of highly conserved 16S rRNA from the metagenome and uses them as seeds to assign the remaining

reads using composition-based clustering. The caveat with SOMs is that it was shown to work well only on DNA fragments that are longer than 8 kbp and lose much accuracy for reads with length below 1 kbp. All the above supervised methods depend on the availability of reference data for training. A metagenomic dataset may, however, contain reads from unexplored phyla which cannot be labeled into one of the existing classes. The accuracy of these methods on dataset containing reads from unknown species is yet to be demonstrated. LikelyBin is an unsupervised method that clusters metagenomic sequences via a Monte Carlo Markov Chain approach [17]. The method was tested on samples that were sufficiently divergent according to derived criteria. Scimm is a recently developed model-based approach to sequence clustering where interpolated Markov models represent clusters and optimization is performed using a variant of the k-means algorithm [18]. In the results section, we compare the accuracy of our proposed method with LikelyBin and Scimm on datasets of different divergences. Abundance Bin can be used to classify reads from species with different abundance levels [19]. However, if it is known a priori that the reads differ widely in their abundances, then we recommend using Abundance Bin over other binning methods.

3. Methods

One of the most common genome signatures is the frequency of occurrence of words (or oligomers) in a DNA sequence [20]. In our method, we model each cluster, containing reads from a species, as a function of probability distributions of words comprising them. The inherent basis of this method is that the set of reads sequenced from a species have a characteristic genome signature that distinguishes it from reads belonging to other species. The distribution of word counts along a genome can be approximated as a Gaussian for shorter, frequent words and as a Poisson for longer words that are rare [21]. We propose an unsupervised multidimensional Naive Bayes Poisson mixture model and derive an Expectation Maximization algorithm for the same. The corresponding algorithm for Gaussian mixture model can be derived similarly. At times, longer words tend to be more discriminatory than the shorter ones [22]. However, with the increase in the length of words, the dimensionality of the data increases exponentially, while the word counts become sparse. To tackle high dimensionality and sparsity of word counts, we impose a clustering structure on the word counts as well. Such a model is called a two-way mixture model. In essence, the proposed method provides a general statistical framework for associating each read with its species of origin, based on its genome signatures.

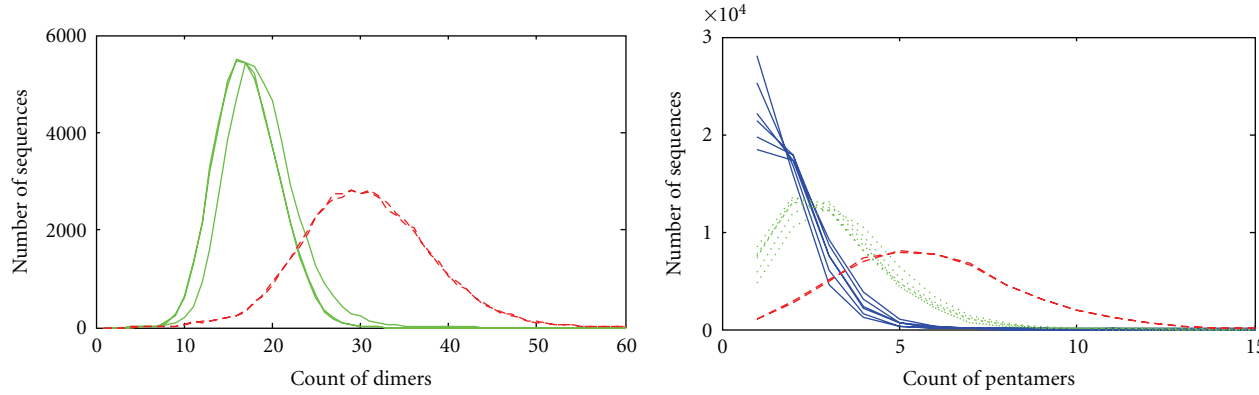
3.1. Need for Multidimensional Word Distribution. A genome signature is a compositional parameter reflecting the relative abundance of different words along the genome. In general, it is similar between closely related species and dissimilar between nonrelated species. Some words that are deemed to be biologically significant are very common in a genome, while others may never be encountered [23]. Composition-based methods use genome signatures to ascertain the origin of the DNA reads. The underlying basis is that the distribution of words in a DNA is specific to each species and undergoes only slight variations along the genome. By establishing the dictionary of words used by a species and their frequency of occurrence, one can point out the basic words of the genome [24].

Literature abounds in methods that study the statistical distribution of the word locations along a sequence and word frequencies [21, 25]. The exact distribution of count of words is known under the hypothesis that the letters are independent (Bernoulli) or under the Markov model. However, in practice, it is extremely time consuming to compute the exact distribution for long sequences or for frequent words. Hence, two kinds of approximations exist. Distribution of word counts along a genome can be approximately modeled as a Gaussian distribution for short words (that are more frequent), or a Poisson distribution for longer words (that are rare) [21].

A metagenomic dataset consists of reads from different species. The reads sampled along a genome of a species will reflect its genome signature. As different words occur with different frequencies along the genome, each word follows its own distribution. Thus, reads belonging to a species can be modeled as a multidimensional distribution of words (one dimension for each word) comprising them. Figure 1 illustrates the distribution of dimer and pentamer counts across reads sampled from the genome of *Haemophilus influenzae* (for the purpose of clarity, only a few distributions are shown). We see that count of each dimer (a short word) across the reads tends to a Gaussian distribution with a different mean and standard deviation and that of a pentamer tends to a Poisson distribution. Hence, the problem of clustering metagenomic reads can be cast as a multidimensional mixture of Gaussians (or Poissons for longer words) where distribution of each word is modeled as a Gaussian (or Poisson). In other words, this corresponds to the multidimensional Naive Bayes model, where each dimension is modeled as a unimodal Gaussian (or Poisson) distribution. Such a general statistical model takes into account the compositional heterogeneity of words along the genome.

3.2. Multispecies Multidimensional Mixture of Distributions. In this paper, we formulate an unsupervised multidimensional Poisson mixture model for clustering reads within a metagenome by their species of origin. We propose to model the reads from a species as a multidimensional distribution of the words comprising them. Therefore, each cluster is represented by the distribution of word counts within the species. The multidimensional model for Gaussian mixtures can be derived analogously. We present the results for both the models.

Mixture models cover the data well, that is, dominant patterns in the data are captured by the component distributions. They allow better approximations of the true distributions, and their parameters are relatively easy to estimate [26]. An additional advantage of using generative models



(a) Distribution of dimers across reads

(b) Distribution of pentamers across reads

Figure 1: Distribution of dimers and pentamers across 50,000 reads sampled from the genome of *Haemophilus influenzae* (only a few distributions are shown). (a) Distribution of dimers tends to Gaussian, two groups can be observed. (b) Distribution of pentamers tends to Poisson, three groups are seen.

is that they are flexible and can handle a large number of classes. For instance, a mixture of Poissons can be multimodal, while a Poisson distribution is always unimodal.

We begin with a metagenome, $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$, containing N reads from M species. Let α_m be the proportion of species m in the dataset, with $\sum_{m=1}^{M} \alpha_m = 1$. We assume that $\mathbf{X}$ is observed and is governed by some density function $p(\mathbf{X} \mid \Theta)$ with parameter Θ. Our goal is to cluster the reads by their species of origin, based on the frequency of words that appear in the reads. For every species m, we want to determine α_m, its proportion in the dataset, and Θ, the parameter governing the distribution of words within the reads. Let $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}$ be the cluster labels. We assume that $y_i = m$ for $m \in 1, \ldots M$ if the ith read belongs to the mth species. Also, $p(y_i = m) = \alpha_m$. Cluster label $\mathbf{Y}$ is unknown. We call $(\mathbf{X}, \mathbf{Y})$ the complete dataset.

For a word of length l, we obtain $p = 4^l$ different words (combinations of A, C, T, G), denoted by $W = \{w_1, w_2, \ldots, w_p\}$. Each read $\mathbf{x_i}$ is represented by a p-dimensional feature vector, $\mathbf{x_i} = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$, where x_{ij} is the count of word w_j in read $\mathbf{x_i}$. We model the distribution of words within every species m by a multidimensional Poisson distribution, say $\boldsymbol{\lambda_m} = \{\lambda_{m1}, \lambda_{m2}, \ldots, \lambda_{mp}\}$. That is, given that read $\mathbf{x_i}$ belongs to species m, the distribution of each word w_j is Poisson with parameter λ_{mj}, where $m = 1, 2, \ldots, M$ and $j = 1, 2, \ldots, p$,

$$p\left(w_j \mid \lambda_{mj}\right) = \phi\left(w_j \mid \lambda_{mj}\right) = \frac{e^{-\lambda_{mj}} \lambda_{mj}^{x_{ij}}}{x_{ij}!}. \tag{1}$$

We assume independence between features of read vector. The probability of a read $\mathbf{x_i}$, given it belongs to species m, is,

$$p(\mathbf{x_i} \mid y_i = m, \Theta) = p(\mathbf{x_i} \mid \lambda_m) = \prod_{j=1}^{p} \phi\left(x_{ij} \mid \lambda_{mj}\right). \tag{2}$$

At first glance, it might seem imprudent to represent a read as a collection of words comprising it, because it leads to the loss in information about the sequencing read. Strictly speaking, even if the sequence of bases in a DNA is independently and identically distributed, distribution of word occurrences are not independent, due to overlaps [21]. Bayesian networks or belief networks can be used to represent the conditional dependencies between the words comprising the reads [27]. Although, in practice, methods for exact inference in Bayesian networks are often computationally expensive. An attractive alternative to Bayesian networks is the Naive Bayes algorithm that assumes independence between the different features of the read. This assumption makes the otherwise complicated problem tractable. Naive Bayes is known to perform well on complex models and takes time that is linear in the number of components. In addition, lost information can be restored at later stages. In this paper, we have presented the formulation of mixture models with the assumption that the different features (word counts) of the read are independent of each other. We outline the Expectation Maximization (EM) algorithm below.

3.3. Parameter Estimation. To initialize the estimation algorithm, we randomly assign each read to a cluster m. The posterior probability $q_{i,m}$ is set to 1, if read i is assigned to cluster m and 0 otherwise. With the initial posterior probabilities, a Maximization step (M-step) is derived to obtain the initial parameters. The EM iterations then follow as follows.

Expectation Step. We estimate the posterior probability $q_{i,m}$ of read $\mathbf{x_i}$ belonging to species m. By Bayes theorem, we have

$$\begin{aligned} p(y_i = m \mid \mathbf{x_i}, \Theta) &= \frac{\alpha_m \cdot p(\mathbf{x_i} \lambda_m)}{\sum_{k=1}^{M} \alpha_k \cdot p(\mathbf{x_i} \lambda_k)} = q_{i,m}, \\ q_{i,m} &\propto \alpha_m \cdot \prod_{j=1}^{p} \phi\left(x_{ij} \mid \lambda_{mj}\right) \text{ subject to } \sum_{m=1}^{M} q_{i,m} = 1. \end{aligned} \tag{3}$$

Maximization Step. The M-step uses $q_{i,m}$ to compute the expectation of complete data log likelihood,

$$\begin{aligned} Q\left(\Theta^{(t+1)}, \Theta^{(t)}\right) &= E_{p(Y|X,\Theta)}[\log\ p(\mathbf{X}, \mathbf{Y} \mid \Theta)] \\ &= \sum_{m=1}^{M}\sum_{i=1}^{N} p\left(y_i = m \mid \mathbf{x_i}, \Theta^{(t)}\right) \\ &\quad \cdot \log\left(p\left(\mathbf{x_i}, y_i = m \mid \Theta^{(t+1)}\right)\right) \\ &= \sum_{m=1}^{M}\sum_{i=1}^{N} (q_{i,m} \cdot \log(\alpha_m \cdot p(\mathbf{x_i} \mid \lambda_m))). \end{aligned} \tag{4}$$

We also take into account the constraint, which requires that α_m's sum to 1 by adding a Lagrange multiplier:

$$\begin{aligned} Q\left(\Theta^{(t+1)}, \Theta^{(t)}\right) &= \sum_{m=1}^{M}\sum_{i=1}^{N} (q_{i,m} \cdot \log(\alpha_m \cdot p(\mathbf{x_i} \mid \lambda_m))) \\ &\quad + \beta\left(\sum_{m=1}^{M} \alpha_m - 1\right). \end{aligned} \tag{5}$$

We maximize the above expression with respect to the parameters, $\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta^{(t+1)}, \Theta^{(t)})$, and update the parameters,

$$\alpha_m^{(t+1)} = \frac{\sum_{i=1}^{N} q_{i,m}}{N}, \qquad \lambda_{mj}^{(t+1)} = \frac{\sum_{i=1}^{N} q_{i,m} \cdot x_{ij}}{\sum_{i=1}^{N} q_{i,m}}. \tag{6}$$

Finally, these two steps are repeated as necessary. Each iteration is guaranteed to increase the log-likelihood, and the algorithm is guaranteed to converge to a local maximum of the likelihood function.

3.4. Word Grouping. Higher-order words are known to be more discriminative than shorter ones [22]. With the increase in length of the word, there are two major consequences that need to be addressed. Firstly, the distribution of words tends to Poisson and not Gaussian (by law of rare numbers), see Figure 1. Secondly, the length of the read vector grows exponentially (e.g., for $l = 10$, $4^l \approx 10^6$). With increase in dimensions, many words will tend to have similar distributions and hence can be clustered together into a "word group." At the same time, the number of distinct words in any read is usually substantially smaller than the number of dimensions. That is, the feature matrix becomes high dimensional and sparse. Hence, the model may fail to predict the true feature distribution of different components. Therefore, dimension reduction becomes necessary before estimating the components in the model. However, reduction of the number of words using feature selection cannot be too aggressive, otherwise the clustering accuracy will suffer.

In this paper, we handle the above challenge by "word grouping." A supervised two-way Poisson mixture model with word grouping was originally proposed by Li and Zha for simultaneous document classification [28]. Such a two-way clustering involves simultaneous clustering of reads as well as of words. The clusters means are regularized by dividing the words into groups and constraining the parameters for the words within the same group to be identical. The grouping of the words is not predetermined but optimized as part of the model estimation. This implies that, for every group, only one statistic for all the words in this group is needed to cluster reads. For instance, in Figure 1, we observe the distribution of pentamers falls into three distinct group. Therefore, words following similar distributions can be clustered together into a "word group."

We extend our formulation to an equivalent two-way unsupervised Poisson mixture model in order to simultaneously cluster word features and classify reads and derive an Expectation Maximization algorithm to estimate its parameters. Figure 2 depicts the paradigm for two-way mixture model of reads. Note that we make a distinction on the use of "cluster" to refer to binning of reads belonging to the same species and "group" to refer to binning of words within read in a cluster.

Recall that the genome signature is similar between closely related species and dissimilar between nonrelated species. The parameter constraint implies that words have the same distribution within each cluster. Therefore, we can assume that, within each cluster, words in different reads have equal Poisson parameters, while, for reads in different clusters, words may follow different Poisson distributions. For simplicity, we assume that all clusters have the same number of word groups. It is trivial to extend to the case where different clusters may have different number of word groups [29].

Let $l \in 1, \ldots, L$ denote the word groups. We define a group assignment function $c(m, j) \in 1, 2, \ldots, L$, which denotes the group to which word w_j belongs in class m. Words in the same word group will have the identical parameters, that is, $\lambda_{mk} = \lambda_{mj} = \theta_{m,l}$, if $c(m, k) = c(m, j)$. The group assignments of the words vary from cluster to cluster. Let the number of words in group l of class m be η_{ml}. The likelihood of $\mathbf{x_i}$ is now

$$p(\mathbf{x_i} \mid \lambda_m) = \prod_{j=1}^{p} p\left(x_{ij} \mid \lambda_{mj}\right) = \prod_{j=1}^{p} p\left(x_{ij} \mid \theta_{m,c(m,j)}\right). \tag{7}$$

Now, we can perform clustering using no more than ML dimensions. Word grouping leads to dimension reduction in this precise sense.

We can derive an EM algorithm similar to the one outlined above to estimate the Poisson parameters $\theta_{m,l}$ where $m \in 1, \ldots, M$, $l \in 1, \ldots, L$, the group assignment function $c(m, j) \in 1, \ldots, L$, where $m \in 1, \ldots, M$, $j \in 1, \ldots, p$ and the prior mixture components α_m, for $m \in 1, \ldots, M$. We initialize by setting each value of the group assignment function $c(m, j)$ randomly to a number in $1, \ldots, L$. We start with the same word group partition for all the clusters, that is, $c(m, j)$s are initially identical over m. We update the parameters as follows:

$$\begin{aligned} \alpha_m^{(t+1)} &= \frac{\sum_{i=1}^{N} q_{i,m}}{N}, \\ \theta_{m,l}^{(t+1)} &= \frac{\sum_{i=1}^{N} q_{i,m} \cdot \sum_{j \in 1} x_{ij}}{\eta_{ml} \sum_{i=1}^{N} q_{i,m}}, \quad \text{where } c(m, j) = l. \end{aligned} \tag{8}$$

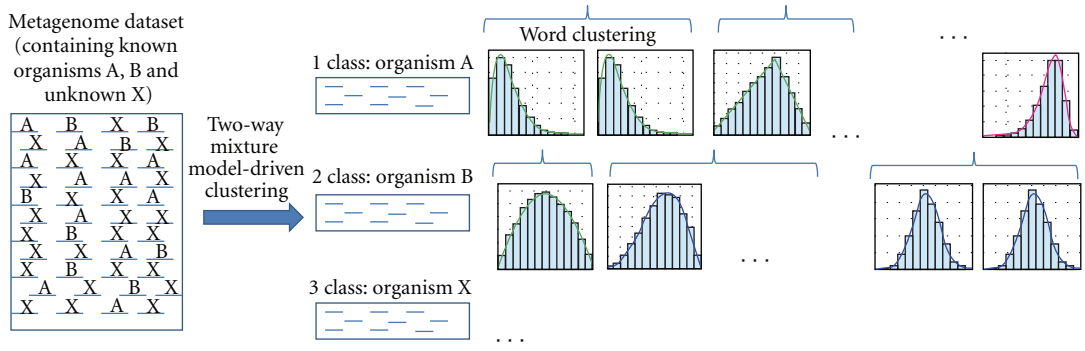


Figure 2: Illustration of a two-way Poisson mixture model for metagenomic data. Each cluster represents a species and is modeled as a distribution of words comprising it. Each word follows a different distribution. However, not all words in a class have significantly different parameters. Therefore, the words can be divided into groups and words within the same group can be constrained to have identical parameters.

Once $\theta_{ml}^{(t+1)}$ is fixed, the word cluster index $c^{(t+1)}(m, j)$ can be found by doing a linear search over all components:

$$c^{(t+1)}(m,j) = \arg\max_{l} \sum_{i=1} q_{i,m}\left(x_{ij}\log\theta_{m,l}^{(t+1)} - \theta_{m,l}^{(t+1)}\right). \quad (9)$$

3.5. Naive Bayes Mixture of Multinomials.

Theorem 1. *If $(X_1, X_2, \ldots, X_p)$ are independent Poisson variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_p$, respectively, then the conditional distribution of $(X_1, X_2, \ldots, X_p)$ given that $X_1 + X_2 + \cdots + X_p = n$ is multinomial with parameters λ_j/λ, where $\lambda = \sum \lambda_j$, that is, Mult(n, π), where $\pi = (\lambda_1/\lambda, \lambda_2/\lambda, \ldots, \lambda_p/\lambda)$ [30].*

The above theorem implies that the unconditional distribution $(X_1, X_2, \ldots, X_p)$ can be factored into a product of two distributions: a Poisson for the overall total and a multinomial distribution of X, $X \sim \text{Mult}(n, \pi)$. Therefore, the likelihood-based inferences about π are the same whether we regard $X_1, X_2, \ldots, X_p$ as sampled from p independent Poissons or from a single multinomial. Here, n refers to the length of the reads and our interest lies in the proportion of words in the reads. Any estimates, tests, inferences about the proportions will be the same whether we regard n as random or fixed.

We can now derive the Naive Bayes mixture of multinomials as standardized mixture of Poissons. We assume that the distribution of words within the reads of a species is governed by the parameters of a multinomial distribution $\boldsymbol{\Theta} = (\theta_{\mathbf{1}}, \theta_{\mathbf{2}}, \ldots, \theta_{\mathbf{m}})$, where each $\theta_{\mathbf{m}}$ is the parameter for species m and is given by $\theta_{\mathbf{m}} = (\theta_{m1}, \theta_{m2}, \ldots, \theta_{mp})$. Therefore, the likelihood of the data will be

$$P\left(\mathbf{x_i} \mid y_i = m\right) = P(\mathbf{x_i} \mid \theta_m) = \frac{n_i!}{\prod_{j=1}^{p} x_{ij}!} \prod_{j=1}^{p} \theta_{mj}^{x_{ij}}, \quad (10)$$

The sum of the probabilities satisfies the constraint $\sum_{j=1}^{p} \theta_{mj} = 1$. The EM algorithm for Naive Bayes mixture of Multinomials can be derived similarly, and we only give the final set of equations:

$$\begin{aligned} \alpha_m &= \frac{\sum_{i=1}^{N} q_{i,m}}{N}, \\ \theta_{mj} &= \frac{\sum_{i=1}^{N} q_{i,m} \cdot x_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{p} q_{i,m} x_{ij}} \\ &= \frac{\sum_{i=1}^{N} q_{i,m} \cdot x_{ij}}{\sum_{i=1}^{N} q_{i,m} n_i}. \end{aligned} \quad (11)$$

If we assume the length of each read to be a constant n, we get the same results as that with Poisson distribution; hence, the two distributions are equivalent in modeling the distribution of words within reads of a species. Also, since the multinomial distribution is single distribution, we do not perform a two-way dimension reduction on it.

4. Results

4.1. Simulated Metagenomes. The algorithm has been implemented in Matlab and C. The space and time complexity scale linearly with the number of reads and species. Space complexity scales quadratically with the number of dimensions in the search space. Our method converged for all the cases we tested and was robust to the choice of initial conditions.

Metagenomics being a relatively new field lacks standard datasets for the purpose of testing clustering algorithms [31]. As the "true solution" for sequence data generated from most metagenomic studies is still unknown, we focused on synthetic datasets for benchmarking. We also apply our method to the actual Acid Mine Drainage dataset to identify the dominant species. In order to test the accuracy of our proposed method, we used MetaSim to simulate synthetic metagenomes [6]. MetaSim takes as input the sequencing technology to be used (Sanger, 454, Exact), a set of known genomes, length of the reads, and an abundance/coverage profile which determines the relative abundance of each genome in the simulated dataset. The genomes used for

generating the synthetic metagenomes were downloaded from National Center of Biotechnology Information (NCBI). We generated datasets with reads of lengths between 50 and 1000 bp and various abundance ratios. In the first part of this section, we demonstrate the performance of the multidimensional Gaussian mixture model on several datasets. A default word length of 2 is used. Additionally, as the number of dimensions is relatively small, we do not perform word grouping. Next, we describe the results using the two-way Poisson mixture model with a word length of 5. The method has been implemented for word lengths from 2 to 9. In order to calculate the clustering accuracy, we assign each cluster to the source species that is most frequent in the cluster. Accuracy is given by the percentage of total correct read assignments.

The number of species in each dataset is supplied as an input. Determining the number of clusters from a statistical perspective is a difficult problem and has been addressed by [32]. Previously, 16 s/18 s rDNA have been used for phylotyping and assessing species diversity using a rarefaction curve [33]. Tools such as MetaPhyler and TreePhyler can be used for making an educated guess of the number of species [34, 35]. Estimating species diversity is still an active area of research, and we do not address it in this paper.

Experiments in the 1960s and 1970s have shown that the dinucleotide relative abundance in a genome is a remarkably stable property [36, 37]. Closely related organisms display more similar dinucleotide composition than do distant organisms [20]. In [38], the authors proposed a measure of intergenomic difference between two sequences f and g, called the average dinucleotide relative abundance,

$$\delta^*(f,g) = \frac{1}{16}\sum_{X,Y} \left|\rho^*_{XY}(f) - \rho^*_{XY}(g)\right|, \qquad (12)$$

where $\rho^*_{XY}(f) = f^*_{XY}/f^*_Y f^*_Y$ and f^*_X denotes the frequency of X in f. A measure of intergenomic difference was obtained by comparing different genome signatures. In order to assess the robustness of our method, we test it across datasets representative of δ^* values ranging from 34 to 340. In general, lower δ^* values correspond to "closely related species" and higher values correspond to "distant species."

In Figure 3, we plot the performance of our proposed multidimensional Poisson model over 450 datasets with δ^* values ranging from 34 to 340. We observed a positive correlation between the intergenomic difference and the accuracy of our method, as also noted in [17]. The initial increase in the accuracy with word length is justified by the increased discriminative power of higher order words. However, any further increase in word length has to be accompanied by dimension reduction, otherwise owing to the high dimensional and sparse nature of feature matrix, the accuracy begins to drop.

In Figure 4, we compare the accuracy of our proposed multidimensional Gaussian model with two other unsupervised composition-based methods LikelyBin [17] and Scimm [18] on several datasets. Default parameters are used for these algorithms. We varied the read length between 200 to 500 bp, δ^* values from 60 to 300, and the abundance ratio up to 1 : 5. Note that the distribution of dimers tends to a Gaussian. As the number of dimensions is relatively small ($4^2 = 16$), the algorithm performs well without word grouping. Our method clearly outperforms LikelyBin and performs as well or better than Scimm on most instances. Another point worth noting from the figure is that our method's error rate is bounded by 10% for datasets with read length as short as 200 bp.

Table 1: Performance of Gaussian mixture model (without word grouping) on datasets containing more than 2 species, at various abundances on reads of length 500 bp. AR stands for abundance ratio.

Species	AR	Number reads	Accuracy (%)
T. thermophilis	1	50000	
A. vinelandii	3		87.51
N. meningitidis	2		
E. coli 536	1	50000	
S. acidocaldarius	2		97.01
H. salinarum R1	2		
C. jejuni RM1221	3	60000	
H. salinarium R1	2		96.61
E. coli	1		
P. horikoshii OT3	3		
S. erythraea	1	60000	
M. thermoautotrophicum	1		90.28
B. burgdorferi ZS7	1		
E. coli 536	1		
B. burgdorferi ZS7	1	75000	
C. jejuni RM1221	1		85.04
E. coli 536	1		
H. salinarum R1	1		
P. horikoshii OT3	1		

We analyzed the accuracy and applicability of our method on binning reads from low complexity communities, containing 3–5 species (see Table 1). With the increase in number of species, there was a slight degradation in performance, though the accuracy was consistently above 85%. This is in agreement with the results from the 2 species dataset, considering that the total coverage of each species is much lower in a multispecies dataset (Reads from *B. Burgdorferi* form only 6% of the 5th dataset).

Next, we evaluated the robustness of our method to changes in the abundance ratio between species as well as the length of the reads. We simulated three sets of metagenomes with two species each at different abundance ratios. We varied the abundance ratio from 10 : 1 to 1 : 10 in stages for the two species. From Figure 5, we note that there was only a slight drop in performance for extreme abundance ratios. Therefore, the proposed method is suited for binning relatively rare species as well. It is noteworthy to point out that estimates are good at all abundances. In order to test the usefulness of the method for analyzing data produced by

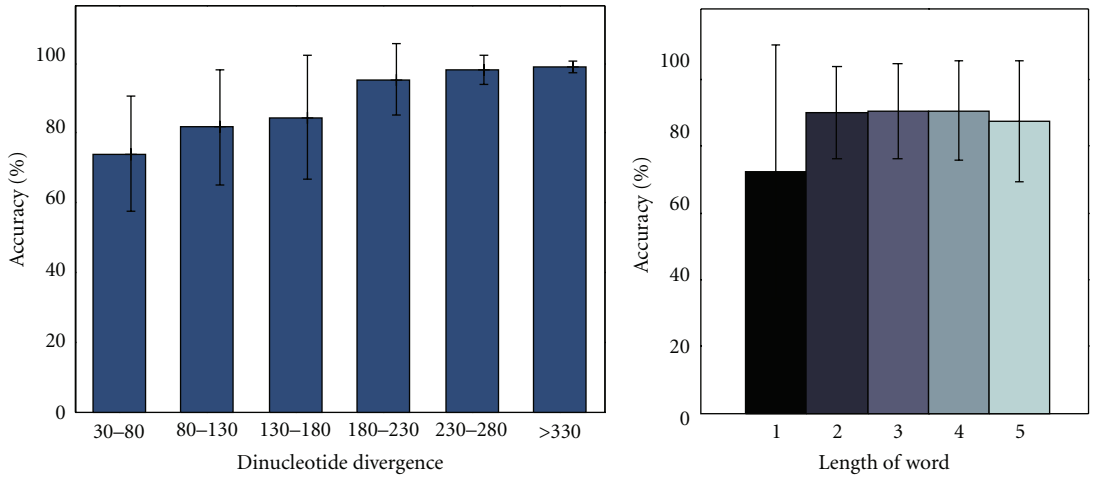


Figure 3: Performance of Poisson Mixture Model at (a) different coverage ratios in a 2-species datasets. Datasets with low δ^* values (100–150) were chosen (b) different reads lengths (50–1000 bp).

the current NGS technologies (especially Solexa and SOLiD) that generate short reads, we tested three datasets of varying δ^* values for read lengths between 50 and 1000 bp. With the decrease in read length from 1000 to 50 bp, the drop in accuracy of our method is bounded by 15%.

Recall that with the increase in the length of the words and the simultaneous increase in the number of dimensions, the distribution of the words tends to a Poisson and word grouping becomes necessary. In this section, we present the clustering results obtained by estimating the two-way Poisson mixture model with different number of word groups L. We observed the variation in classification accuracy to be more prominent for lower values of L. Therefore, in Table 2, we report the results for values of $L < 50$ for a 2 species dataset. If word grouping is not performed, then clustering based on mixture model is essentially the Naive Bayes algorithm with each dimension modeled by a Poisson distribution (last column of Table 2). From the results, we can infer that word grouping resulted in considerable increase in accuracy compared to the Naive Bayes algorithm. That is, the characteristic vectors are of a much lower dimension with $L \ll p$. Also, a high clustering accuracy can be achieved using no more than ML dimensions, significantly smaller than the original dimension, 1024. Note that it is difficult to know a priori, the exact value of L that yields the best clustering. However, among the values we tested, lower values of L provided a higher accuracy.

In Table 3, we compare the performance of our 2-way Poisson mixture model with Gaussian mixture model for datasets with low δ^* values. In real situations, it is difficult to know beforehand, the most discerning order of the word to use. However, from our experiments, we can infer that higher-order word-based models, in general, tend to be more discriminatory than those based on lower order words. If it is known a priori that lower-order words (of length 2-3) are more discriminatory in the dataset, then we recommend using a Gaussian mixture model. For other datasets, we use a Poisson mixture model.

Our method's accuracy in classifying reads from the datasets composed of species across various taxonomic ranks is reported in Table 4 we used the Poisson mixture model without word grouping. The error rates are bounded by 10% on all datasets. We can infer that the accuracy is mostly correlated to the phylogenetic distances between the species. For example, reads from datasets containing species with taxonomic differences at the level of class were classified with a very high accuracy.

4.2. Real Metagenome: Acid Mine Drainage (AMD) Dataset. The ultimate goal of binning methods is to cluster reads in a real metagenome, by their species of origin. Clustering in real situations is error-prone and affects our final estimates of species abundance. Moreover, evaluating clustering methods on real metagenomes can be problematic as the true taxonomic composition of the data is mostly unknown. The accuracy of unsupervised clustering methods decreases with increase in the complexity of metagenomes and for species present at very low abundances. However, the composition of Acid Mine Drainage metagenome has been substantially characterized, and we used this dataset to evaluate the performance of our proposed method [39]. The AMD microbial community is reported to consist of two dominant populations (*Ferroplasma sp. Type II* and *Leptospirillum sp. Group II*) and three other less abundant ones (*Ferroplasma acidarmanus Type I, Leptospirillum sp. Group III,* and *Thermoplasmatales archaeon GpI*). We downloaded the reads, as well as the scaffolds assembled from the reads for the 5 species of the actual AMD dataset from NCBI. Only 58% of the AMD reads can be mapped back to the assembled scaffolds using BLAST [19]. Therefore, in order to compute the accuracy of our method, we simulated a metagenome

Table 2: Performance of Poisson mixture model on datasets for different values of L and word length of 5. Here, N.W.G stands for no word grouping. The maximum accuracy achieved is in bold. Each dataset contains 50,000 reads of length 500 bp.

Species	$L = 5$	$L = 10$	$L = 30$	$L = 50$	N.W.G
B. anthracis CI chromosome, *B. halodurans* C-125	90.61	**91.53**	50.31	91.2	50.32
H. pylori 26695, *S. pneumoniae* 70585	98.6	**98.79**	98. 73	98.71	98.76
B. subtilis subsp. spizizenii str., L. lactis subsp.	89.96	90.34	**90.62**	90.53	50.47

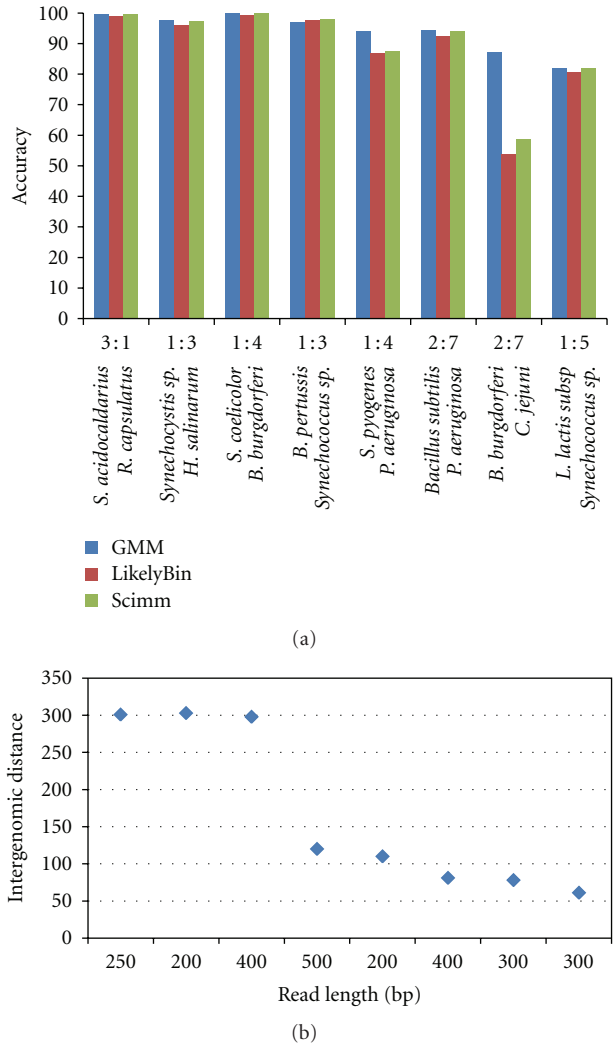


Figure 4: GMM stands for Gaussian mixture model (without word grouping). (a) compares the performance of the three methods on 8 datasets (X-axis shows the abundance ratio and the species contained in the dataset). (b) plots the δ^* values for the corresponding datasets. The X-axis shows the corresponding read lengths. Here, the δ^* (measured on 50 kb contigs) ranges from 34 to 340.

with reads sampled from the downloaded scaffolds. The simulated AMD dataset consisted of 110,000 reads of average length 732 bp (average read length in the actual AMD dataset) from the 5 species, in the ratio 4 : 4 : 1 : 1 : 1. We characterized the dataset in two stages. Notice that the

Table 3: Comparison of performance of Gaussian mixture model (GMM) with 2-way Poisson mixture model (PMM) for datasets with low δ^* values. Each dataset contains 50,000 reads of length 500 bp.

Species	δ^*	GMM	PMM
M. leprae, P. putida	74	75.25	85.24
B. subtilis, L. lactis	86	86.23	90.62
H. pylori, S. pneumoniae	148	53.48	98.76
H. salinarum, R. sphaeroides	153	94.63	98.51
M. jannaschii, S. aureus	164	50.0	97.75

Table 4: Performance of Poisson mixture model (without word grouping) on datasets across various taxonomic ranks. Each dataset contains 50,000 reads of length 500 bp. AR stands for abundance ratio.

Species	AR	Rank	Accuracy (%)
M. hyopneumoniae, M. mycoides	3 : 2	Genus	95.73
M. avium, M. leprae	3 : 4	Genus	94.22
A. vinelandii, C. japonicus	1 : 1	Family	92.81
M. leprae, S. erythraea	1 : 1	Order	95.58
B. pertussis, N. gonorrhoeae	1 : 2	Class	97.52
A. parvulum, S. erythraea	5 : 1	Class	99.64
R. prowazekii, S. meliloti	3 : 1	Class	99.91

dataset contains reads with two distinct abundance levels. Therefore, we can simplify the problem by first separating the reads into two bins based on their abundance. In the first stage, the reads were grouped into two bins, using Abundance Bin, with a resulting accuracy of 93.3%. The bins corresponding to the abundance levels of 4 and 1 had a cluster purity of 93.2% and 98.2%, respectively. In the next stage, we used the reads from each of the bins output by Abundance Bin, as an input to our proposed 2-way Poisson mixture model, to further classify the reads by their species of origin. We used a word length of 5. Our method clustered the reads from the bin containing dominant species into two clusters corresponding to *Ferroplasma sp. Type II* and *Leptospirillum sp. Group II*, with an accuracy of 96.88% (with $L = 10$). The other bin consisted of very few reads from the remaining three species *Ferroplasma acidarmanus Type I, Leptospirillum sp. Group III*, and *Thermoplasmatales archaeon GpI*. Our method clustered the reads from this bin into three clusters, with an accuracy of 70.34% (with $L = 10$). This decrease in accuracy can be attributed to the low bin count.

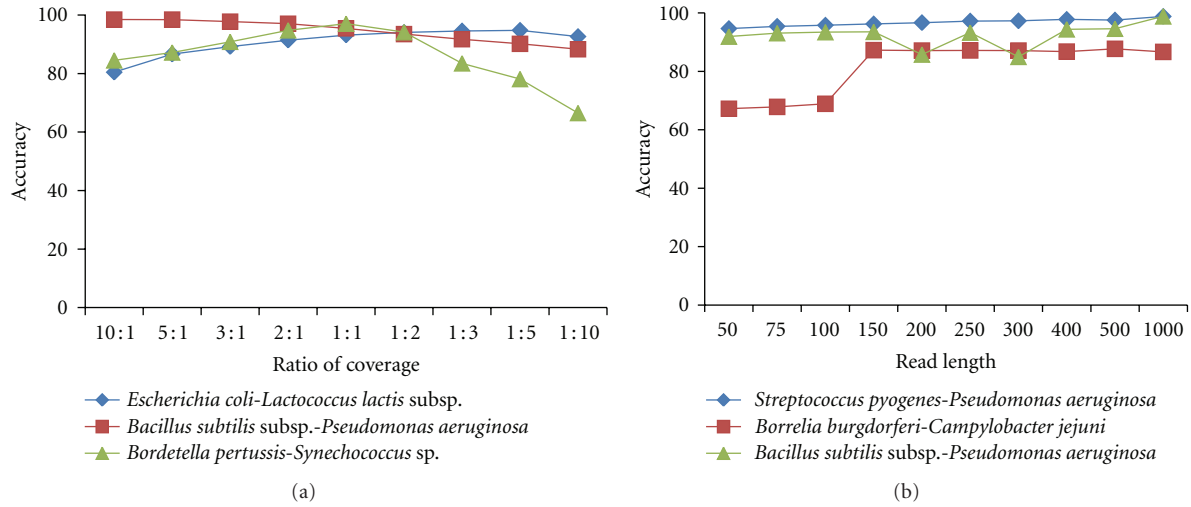


FIGURE 5: Performance of Poisson mixture model at (b) different reads lengths (50–1000 bp), (a) different coverage ratios in a 2-species datasets. Datasets with low δ^* values (100–150) were chosen.

5. Discussion

In this paper, we formulated an unsupervised two-way multispecies, multidimensional mixture model to represent reads from a metagenome. We used the proposed model to cluster metagenomic reads by their species of origin and to characterize the abundance of each species. The distribution of word counts along a genome can be approximated as a Gaussian for shorter, frequent words and as a Poisson for longer words that are rare. Therefore, we use a multidimensional mixture of Gaussians or Poissons to model the reads from each bin. An additional reason to use these distributions is their flexibility, stability, and ease of parameter estimation. Our method is an unsupervised method that does not require any training data. This is critical for success as most metagenomic datasets contain reads from unexplored phyla which cannot be labeled into one of the existing classes. Our probabilistic approach can be used to identify reads which belong to more than one species and occlude the cluster boundaries. Such reads should be further investigated to identify the presence of conserved regions.

Note that our proposed method is primarily a composition-based method that seeks to distinguish between genomes based on their characteristic DNA compositional pattern. Therefore, it cannot distinguish between genomes unless their DNA compositions are sufficiently divergent (see Figure 4, dataset with *B. Burgdorferi* and *C. Jejuni*). It is unlikely that our method will be able to accurately distinguish between strains of the same species. For such datasets, genome signature alone is insufficient for inferring taxonomic relationships reliably. Composition-based methods must be used in conjunction with other similarity-based methods and abundance-based methods to yield better performance.

Note that the two-way Poisson mixture model was originally proposed for classification of documents. In this work, we demonstrate the relevance and applicability of such a general statistical framework for modeling metagenome reads. We have illustrated that the proposed method can accurately classify reads from low to medium complexity datasets into taxon-specific bins, based on genome signatures.

Our framework complements the existing similarity-based and abundance-based methods and hence can be combined with such methods to obtain a better performance. We intend to develop such hybrid methods in the future that can tackle the problem of classifying sequences in complex metagenomic communities.

Acknowledgment

The authors wish to thank Jia Li for her useful insights into the two-way Poisson mixture model problem and the reviewers for their valuable comments and suggestions.

References

[1] K. Chen and L. Pachter, "Bioinformatics for whole-genome shotgun sequencing of microbial communities," *PLoS Computational Biology*, vol. 1, no. 2, article e24, pp. 0106–0112, 2005.

[2] M. S. Rappé and S. J. Giovannoni, "The uncultured microbial majority," *Annual Review of Microbiology*, vol. 57, pp. 369–394, 2003.

[3] M. Pop, "Genome assembly reborn: recent computational challenges," *Briefings in Bioinformatics*, vol. 10, no. 4, pp. 354–366, 2009.

[4] C. K. K. Chan, A. L. Hsu, S. K. Halgamuge, and S. L. Tang, "Binning sequences using very sparse labels within a metagenome," *BMC Bioinformatics*, vol. 9, article no. 215, 2008.

[5] S. Nasser, A. Breland, F. Harris, and M. Nicolescu, "A fuzzy classifier to taxonomically group," in *Proceedings of the DNA Fragments within a Metagenome*, vol. New york, NY, USA, pp. 1–6, 2008, Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS '08).

[6] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Research*, vol. 17, no. 3, pp. 377–386, 2007.

[7] A. C. McHardy, H. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos, "Accurate phylogenetic classification of variable-length DNA fragments," *Nature Methods*, vol. 4, no. 1, pp. 63–72, 2007.

[8] A. Brady and S. L. Salzberg, "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models," *Nature Methods*, vol. 6, no. 9, pp. 673–676, 2009.

[9] S. Nasser, A. Breland, F. Harris, and M. Nicolescu, "Metagenome fragment classification using n-mer frequency profiles," in *Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS '08)*, pp. 1–6, New York, NY, USA, 2008.

[10] S. Chatterji, I. Yamazaki, Z. Bai, and J. A. Eisen, "CompostBin: a DNA composition-based algorithm for binning environmental shotgun reads," *Lecture Notes in Computer Science*, vol. 4955, pp. 17–28, 2008.

[11] F. Gori, G. Folino, M. S.M. Jetten, and E. Marchiori, "MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks," *Bioinformatics*, vol. 27, no. 2, pp. 196–203, 2011.

[12] L. Krause, N. N. Diaz, A. Goesmann et al., "Phylogenetic classification of short environmental DNA fragments," *Nucleic Acids Research*, vol. 36, no. 7, pp. 2230–2239, 2008.

[13] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[14] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics*, vol. 5, no. 1, article no. 163, 2004.

[15] M. Bailly-Bechet, A. Danchin, M. Iqbal, M. Marsili, and M. Vergassola, "Codon usage domains over bacterial chromosomes," *PLoS Computational Biology*, vol. 2, no. 4, article e37, pp. 263–275, 2006.

[16] S. D. Bentley and J. Parkhill, "Comparative genomic structure of prokaryotes," *Annual Review of Genetics*, vol. 38, pp. 771–792, 2004.

[17] A. Kislyuk, S. Bhatnagar, J. Dushoff, and J. S. Weitz, "Unsupervised statistical clustering of environmental shotgun sequences," *BMC Bioinformatics*, vol. 10, article no. 1471, p. 316, 2009.

[18] D. R. Kelley and S. L. Salzberg, "Clustering metagenomic sequences with interpolated Markov models," *BMC Bioinformatics*, vol. 11, no. 1, article no. 544, 2010.

[19] Y.-W. Wu and Y. Ye, "A novel abundance-based algorithm for binning metagenomic sequences using l-tuples," *Lecture Notes in Computer Science*, vol. 6044, pp. 535–549, 2010.

[20] S. Karlin, I. Ladunga, and B. E. Blaisdell, "Heterogeneity of genomes: measures and values," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 26, pp. 12837–12841, 1994.

[21] G. Reinert, S. Schbath, and M. S. Waterman, "Probabilistic and statistical properties of words: an overview," *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 1–46, 2000.

[22] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology*, vol. 6, no. 9, pp. 938–947, 2004.

[23] V. Brendel, J. S. Beckmann, and E. N. Trifonov, "Linguistics of nucleotide sequences: morphology and comparison of vocabularies," *Journal of Biomolecular Structure and Dynamics*, vol. 4, no. 1, pp. 011–021, 1986.

[24] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertit, "Genomic signature: characterization and classification of speies assessed by chaos game representation of sequences," *Molecular Biology and Evolution*, vol. 16, no. 10, pp. 1391–1399, 1999.

[25] S. Robin, F. Rodolphe, and S. Schbath, *DNA, Words and Models: Statistics of Exceptional Words*, Cambridge University Press, 2005.

[26] Y. Song, Z. Zhuang, H. Li et al., "Real-time automatic tag recommendation," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '08)*, pp. 515–522, ACM, New York, NY, USA, 2008.

[27] J. Pearl, *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*, Morgan Kaufmann, 1988.

[28] J. Li and H. Zha, "Two-way Poisson mixture models for simultaneous document classification and word clustering," *Computational Statistics and Data Analysis*, vol. 50, no. 1, pp. 163–180, 2006.

[29] M. Qiao and J. Li, "Two-way Gaussian mixture models for high dimensional classification," *Statistical Analysis and Data Mining*, vol. 3, no. 4, pp. 259–271, 2010.

[30] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, Wiley, 1968.

[31] K. Mavromatis, N. Ivanova, K. Barry et al., "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods," *Nature Methods*, vol. 4, no. 6, pp. 495–500, 2007.

[32] R. Tibshirani and G. Walther, "Cluster validation by prediction strength," *Journal of Computational and Graphical Statistics*, vol. 14, no. 3, pp. 511–528, 2005.

[33] J. R. Cole, B. Chai, R. J. Farris et al., "The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis," *Nucleic Acids Research*, vol. 33, supplement 1, pp. D294–D296, 2005.

[34] B. Liu, T. Gibbons, M. Ghodsi, and M. Pop, "Metaphyler: taxonomic profiling for metagenomic sequences," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM'10)*, T. Park, S. K.-W. Tsui, L. Chen, M. K. Ng, L. Wong, and X. Hu, Eds., p. 95, IEEE Computer Society, 2010.

[35] F. Schreiber, P. Gumrich, R. Daniel, and P. Meinicke, "Treephyler: fast taxonomic profiling of metagenomes," *Bioinformatics*, vol. 26, no. 7, Article ID btq070, pp. 960–961, 2010.

[36] J. Josse, A. D. Kaiser, and A. Kornberg, "Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid," *The Journal of Biological Chemistry*, vol. 236, pp. 864–875, 1961.

[37] G. J. Russell, P. M. B. Walker, R. A. Elton, and J. H. Subak Sharpe, "Doublet frequency analysis of fractionated vertebrate nuclear DNA," *Journal of Molecular Biology*, vol. 108, no. 1, pp. 1–20, 1976.

[38] A. Campbell, J. Mrázek, and S. Karlin, "Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 16, pp. 9184–9189, 1999.

[39] G. W. Tyson, J. Chapman, P. Hugenholtz et al., "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, no. 6978, pp. 37–43, 2004.