A Two-Way Bayesian Mixture Model for Clustering in Metagenomics

Shruthi Prabhakara and Raj Acharya

Pennsylvania State University, University Park, PA, 16801
{shruthi,acharya}@psu.edu

Abstract. We present a new and efficient Bayesian mixture model based on Poisson and Multinomial distributions for clustering metagenomic reads by their species of origin. We use the relative abundance of different words along a genome to distinguish reads from different species. The distribution of word counts within a genome is accurately represented by a Poisson distribution. The Multinomial mixture model is derived as a standardized Poisson mixture model. The Bayesian network efficiently encodes the conditional dependencies between word counts in a DNA due to overlaps and hence is most consistent with the data. We present a two-way mixture model that captures the high dimensionality and sparsity associated with the data. Our method can cluster reads as short as 50 bps with accuracy over 80%. The Bayesian mixture models clearly outperform their Naive Bayes counterparts on datasets of varying abundances, divergences and read lengths. Our method attains comparable accuracy to that of state-of-art Scimm and converges at least 5 times faster than Scimm for all the cases tested. The reduced time taken, by our method, to obtain accurate results is highly significant and justifies the use of our proposed method to evaluate large metagenome datasets.

Keywords: Clustering, Mixture Modeling, Metagenomics.

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[6]. It has shown tremendous potential to discover and study the vast majority of species that are resistant to cultivation and sequencing by traditional methods. Unlike single genome sequencing, assembly of a metagenome is intractable and is by large, an unsolved mystery.

A crucial step in metagenomics that is not required in single genome assembly, is binning the reads belonging to a species i.e. the need to associate the reads with its source organism. Clustering methods aim to identify the species present in the sample, classify the sequences by their species of origin and quantify the abundance of each of these species. 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.

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 25–36, 2011.

2 Related Work

Current approaches to metagenome clustering can be mainly classified into similarity-based and composition-based methods. The similarity-based approaches align reads to close phylogenetic neighbors and hence depend on the availability of closely related genomes in existing databases[7,11]. 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. On the other hand, composition-based methods rely on the intrinsic features of the reads such as oligomer/word distributions[15,3,5,13,12], codon usage preference[1] and GC composition[2] to ascertain the origin of the 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.

Most of the existing clustering methods are supervised and depend on the availability of reference data for training[15,3,19,5]. A metagenome may however, contain reads from unexplored phyla which cannot be labeled into one of the existing classes. Most clustering methods until now have been relatively inaccurate in classifying short reads. The poor performance on short reads can be attributed to the high dimensionality and sparsity associated with the data[5].

LikelyBin is an unsupervised method that clusters metagenomic sequences via a Monte Carlo Markov Chain approach[13]. Scimm is a recently developed state-of-art model-based approach where interpolated Markov models represent clusters and optimization is performed using a variant of the k-means algorithm (initialized by LikelyBin and CompostBin)[12]. Later, we compare the performance of our proposed method with Scimm.

Mixture models have become popular tools for analyzing biological sequence data. The dominant patterns in the data are captured by its component distributions. Most mixture models assume an underlying normal distribution[19]. However, the distribution of word counts within a genome vary according to a Poisson distribution[17,18]. The Poisson distribution is adequately approximated by a normal distribution for short words with high count. However, when the count is low, a Poisson distribution more accurately represents the data[24]. Figure 1 illustrates the distribution of dimers and pentamers across reads sampled from the genome of Haemophilus Influenzae. Therefore, the problem of clustering metagenomic reads where distribution of each word varies according to a Poisson distribution can be cast as a multivariate mixture of Poissons.

Bayesian and Naive Bayes Models: Bayesian networks have been an active area of research[10]. Bayesian networks can efficiently represent complex probability distributions. It encodes the joint probability distribution of a set of n variables, $\{X_1, X_2, ..., X_n\}$ as a directed acyclic graph and a set of conditional probability distributions (CPDs). The set of parents of X_i are denoted by Pa_i. Each X_i is conditionally dependent only on its parents Pa_i. The joint probability distribution is given by,

$$p(X_1, X_2, ..., X_n|\Theta) = \prod_{i=1}^{n} p(X_i|Pa_i, \Theta) \tag{1}$$

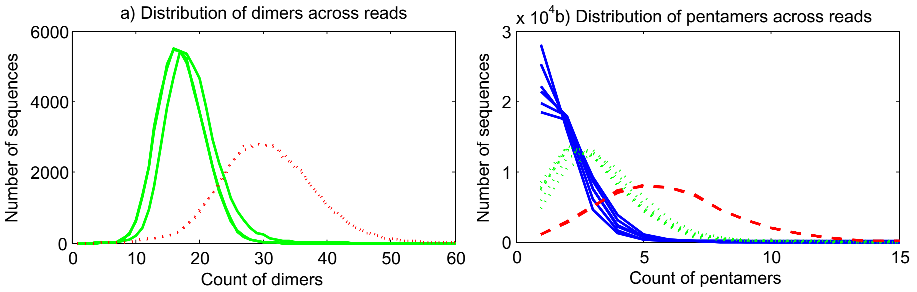

Fig. 1. Distribution of dimers and pentamers across 50,000 reads sampled from the genome of Haemophilus Influenzae(Only a few distributions are shown). a) Distribution of dimers tends to Gaussian and is approximated by a Poisson, two groups can be observed. b) Distribution of pentamers tends to Poisson, three groups are observed.

Typically, even if the sequence of bases in a DNA are independently and identically distributed, distribution of word counts are not independent due to overlaps. Hence, Bayesian networks are ideal for representing sequence data. Though, in practice, methods for exact inference of the structures in Bayesian networks are often computationally expensive. An alternative to Bayesian networks is the Naive Bayes method that assumes independence between the variables. It takes time linear in the number of components. The joint probability distribution is given by,

$$p(X_1, X_2, ..., X_n|\Theta) = \prod_{i=1}^{n} p(X_i|\Theta) \tag{2}$$

Naive Bayes is the simplest Bayesian network that does not represent any variable dependencies. In [20], we described a Naive Bayes mixture of Poissons model to cluster the metagenome reads. We implicitly assume that the variables within a class are independent. The motivation in this paper is to overcome the bottleneck of Naive Bayes by taking into account the conditional dependencies between the word counts within the reads. We focus on developing a tractable Bayesian network for a mixture of Poisson and Multinomial distributions.

3 Methods

We are given a metagenome, $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$, 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 $p(\mathbf{X}|\Theta)$ 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 $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$, be the cluster labels. We assume that $y_i = m$ for $m \in 1, ...M$, if the i^{th} read belongs to the m^{th} species. Also, $p(y_i = m) = \alpha_m$. Cluster label $\mathbf{Y}$ is unknown. We call $(\mathbf{X}, \mathbf{Y})$, the complete dataset.

We use Bayesian networks to represent the conditional dependencies between words. Let read $\mathbf{x}$ be of length n , $\mathbf{x} = (c_1 c_2 ... c_n)$, where each $c_k \in (A, C, T, G)$. We assume that the probability of the read is determined by a set of $p = 4^l$ probabilities corresponding to words of length l.

$$p(\mathbf{x}|\Theta) = p(c_1 c_2 ... c_l) \prod_{k=l+1}^{n} p(c_k | c_{k-l} ... c_{k-1}, \Theta) = \prod_{k=1}^{n} p(c_k | pa_k, \Theta) \tag{3}$$

In a read, any given nucleotide c_k can be preceded by its parents pa_k in the read, where $c_k \in (A, C, T, G)$ and $pa_k \in \{pa_k^1, pa_k^2, .., pa_k^p\}$ denote different word configurations of parents. In the next section, we will formulate the Bayesian mixture of Poissons from first principles. In section 3.2, we present the two-way Bayesian mixture of Poissons that uses “word grouping” to handle high-dimensionality and sparsity associated with the metagenome. In section 3.3, we briefly introduce the Bayesian mixture of Multinomials as standardized Bayesian mixture of Poissons and the corresponding two-way Bayesian mixture of Multinomials.

3.1 Bayesian Mixture of Poissons

We represent each read $\mathbf{x_i}$ by a $4 \times p$ matrix $\mathbb{N}_i = \{N_i(c_k|pa_j) : j = 1, ..., p\}$ and $c_k \in (A, C, T, G)$, where $N_i(c_k|pa_j)$ is the count of the number of occurrences of parent word pa_j followed by nucleotide c_k in read $\mathbf{x_i}$. The distribution of words within the reads of a species follow the parameters of a Poisson distribution, $\mathbf{\Theta} = (\lambda_\mathbf{1}, \lambda_\mathbf{2}, ..., \lambda_\mathbf{m})$, where $\lambda_\mathbf{m} = ((\lambda_{m,c_k|pa_j})_{\forall c_k})_{\forall pa_j}$, i.e., each local species distribution is a collection of Poisson distributions, one for every configuration pa_j of parents and c_k, and has the same parameters across reads of a species.

$$\begin{aligned} &\mathbf{\Theta} = \{\lambda_\mathbf{m} : \forall m \in 1, .., M\} \\ &\lambda_\mathbf{m} = \{\lambda_{m,c_k|pa_j} : \forall c_k \in (A, C, T, G) \text{ and } \forall pa_j \in \{pa_j^1, pa_j^2, .., pa_j^p\}\} \end{aligned} \tag{4}$$

Therefore, the likelihood of the data will be,

$$\begin{aligned} p(\mathbf{x_i}|y_i = m) = p(\mathbf{x_i}|\lambda_\mathbf{m}) &= \prod_{pa_j} \prod_{c_k} p(N_i(c_k|pa_j)|\lambda_{m,c_k|pa_j}) \\ &= \prod_{pa_j} \prod_{c_k} \frac{\lambda_{m,c_k|pa_j}^{N_i(c_k|pa_j)} e^{-\lambda_{m,c_k|pa_j}}}{N_i(c_k|pa_j)!} \end{aligned} \tag{5}$$

EM Algorithm: We use Expectation-Maximization (EM) algorithm to infer the parameters [8]. In the expectation step, use the current parameter estimate $\Theta^{(i-1)}$ to find the posterior probability $p(y_i = m|\mathbf{x_i}, \Theta^{(i-1)})$ or $q_{i,m}$.

$$q_{i,m} \propto \alpha_m . \prod_{pa_j} \prod_{c_k} p(N_i(c_k|pa_j)|\lambda_{m,c_k|pa_j}) \text{ subject to } \sum_{m=1}^{M} q_{i,m} = 1$$

In the maximization step, determine the expectation of the complete-data log likelihood.

$$Q(\Theta, \Theta^{(i-1)}) = \sum_{m=1}^{M} \sum_{i=1}^{N} q_{i,m} \Bigg(\log(\alpha_m) + \sum_{pa_j} \sum_{c_k} (N_i(c_k|pa_j) \log \lambda_{m,c_k|pa_j} - \lambda_{m,c_k|pa_j}) \Bigg) \tag{6}$$

subject to the constraint,$\sum_{m=1}^{M} \alpha_m = 1$. The maximum likelihood estimates for the Bayesian mixture of Poissons are,

$$\alpha_m = \frac{\sum_{i=1}^{N} q_{i,m}}{N} \ , \ \lambda_{m,c_k|pa_j} = \frac{\sum_{i=1}^{N} q_{i,m}.N_i(c_k|pa_j)}{\sum_{i=1}^{N} q_{i,m}} \tag{7}$$

To initialize the EM 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. We then proceed with the M-step. Each iteration is guaranteed to increase the log-likelihood and the algorithm is guaranteed to converge to a local maximum of the likelihood function.

3.2 Two-Way Bayesian Mixture of Poissons

Higher order words are known to be more discriminative than shorter ones[21]. However, with the increase in the length of words, the length of the read vector grows exponentially (e.g, for $l = 10, 4^l \approx 10^6$). Moreover, many words will tend to similar distributions and hence, can be clustered together into a "word group". The feature matrix becomes high-dimensional and sparse. Hence, the model may fail to predict the true 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". This idea was first explored by Li *et al.* for a Naive Bayes mixture of Poisson distributions [14]. They called such a model a two-way mixture model, reflecting the observation that the mixture clusters induce a partition of the reads as well as of words. The cluster 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 pre-determined, 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, the distributions of pentamers falls into three distinct groups. Thus, words following similar distributions can be clustered together into a "word group". Note that we make a distinction on the use of "cluster" for binning of reads within the same species and "group" for binning of words within a cluster. For simplicity, we assume that all clusters have the same number of word groups.

In the Bayesian mixture of Poissons, we group the set of Poisson parameters corresponding to each parent into its Poisson vector. Therefore, we have p different Poisson vectors corresponding to p configurations of parents. We divide the parents into groups and constrain the Poisson vector distributions corresponding to parents within the same group to have identical parameters. Let L be the number of groups within each cluster. Let $c(m, pa_j) \in 1, 2, ..., L$ denote the group that pa_j belongs to in class m. All parents in group l, have Poisson parameter $\lambda_{m,c_k|l}$. Let the number of parents in group l of class m be η_{ml}.

$$p(\mathbf{x_i}|\lambda_{\mathbf{m}}) = \prod_{pa_j}\prod_{c_k} \frac{(\lambda_{m,c_k|l}^{N_i(c_k|pa_j)} e^{-\lambda_{m,c_k|l}})}{N_i(c_k|pa_j)!} \text{ where } c(m, pa_j) = l \tag{8}$$

Now, we can perform clustering using no more than order of 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 parameters.

$$\alpha_m = \frac{\sum_{i=1}^{N} q_{i,m}}{N}\ ,\ \lambda_{m,c_k|pa_j} = \frac{\sum_{i=1}^{N} q_{i,m}.\sum_{pa_j\in l} N_i(c_k|pa_j)}{\eta_{ml}\sum_{i=1}^{N} q_{i,m}} \tag{9}$$

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

$$c(m, pa_j) = \arg\max_{l} \sum_{i=1} q_{i,m} \sum_{c_k} (x_{ij} \log(\lambda_{m,c_k|l}) - \lambda_{m,c_k|l})) \tag{10}$$

3.3 Bayesian Mixture of Multinomials

Theorem: If $(X_1, X_2, .., X_p)$ are independent Poisson variables with parameters, $\lambda_1, \lambda_2, .., \lambda_p$ respectively, then the conditional distribution of $(X_1, X_2, .., X_p)$ given that $X_1 + X_2 + ... + X_p = n$ is multinomial with parameters λ_j/λ, where $\lambda = \sum \lambda_j$, i.e. $Mult(n, \pi)$, where $\pi = (\lambda_1/\lambda, \lambda_2/\lambda, ..., \lambda_p/\lambda)$[9].

The above theorem implies that the unconditional distribution $(X_1, X_2, ..., X_p)$ can be factored into a product of two distributions: a Poisson for the overall total, and a multinomial distribution of X, $X \sim Mult(n, \pi)$. Therefore, the likelihood based inferences about π are the same whether we regard $X_1, X_2, .., X_p$ 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 Bayesian mixture of Multinomials as standardized Bayesian 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 $\mathbf{\Theta} = (\theta_{\mathbf{1}}, \theta_{\mathbf{2}}, ..., \theta_{\mathbf{m}})$. Let $P_m(c_k|pa_j) = \theta_{m,c_k|pa_j}$. The sum of CPDs is well-defined, $\sum_{c_k\in(A,C,T,G)} \theta_{m,c_k|pa_j} = 1\ \forall m$ and $\forall pa_j$. Each local species distribution is a collection of multinomial distributions, one for each configuration of pa_j. $\theta_{\mathbf{m}} = \{((\theta_{m,c_k|pa_j})_{\forall c_k})_{\forall pa_j}\}$. Therefore, within every species m, for each configuration pa_j of parents, we get an independent multinomial problem, $Mult(\theta_{\mathbf{m,c|pa_j}}) = (\theta_{m,c_k|pa_j})\forall c_k$, that has the same parameters across reads of a species.

$$\begin{aligned}
&\boldsymbol{\Theta} = \{\theta_{\mathbf{m}} : \forall m \in 1, .., M\}\\
&\theta_{\mathbf{m}} = \{\theta_{\mathbf{m,c|pa_j}} : \forall pa_j \in \{pa_j^1, pa_j^2, .., pa_j^p\}\}\\
&Mult(\theta_{\mathbf{m,c|pa_j}}) = \theta_{\mathbf{m,c|pa_j}} = \{\theta_{m,c_k|pa_j} : \forall c_k \in (A, C, T, G)\}
\end{aligned} \tag{11}$$

Therefore, the likelihood of the data will be,

$$p(\mathbf{x_i}|y_i = m) = p(\mathbf{x_i}|\theta_{\mathbf{m}}) = \prod_{pa_j} \prod_{c_k \in (A,C,T,G)} \theta_{m,c_k|pa_j}^{N_i(c_k|pa_j)} \tag{12}$$

The EM algorithm for Bayesian mixture of Multinomials and its corresponding two-way reduction can be derived similarly and we do not discuss it further.

4 Results

Datasets: Metagenomics being a relatively new field, lacks standard datasets for the purpose of testing clustering algorithms. As the "true solution" for sequence data generated from most metagenomic studies is still unknown, we focus on synthetic datasets for benchmarking. We use Metasim to simulate synthetic metagenomes[11]. It takes as input the sequencing technology to be used, a set of known genomes, length of the reads and a profile that determines the relative abundance of each genome in the dataset. We generated over 450 datasets with read lengths between 50 and 1000 bps and various abundance ratios.

The algorithms were implemented in Matlab. The space and time complexity scale linearly with the number of reads and species and quadratically with the number of dimensions in the search space. Our methods converged for all the cases we tested and was robust to the choice of initial conditions.

In order assess the robustness of our method, we ranked the 2-species datasets by a measure of intergenomic difference between sequences f and g, called the average dinucleotide relative abundance [4].

$$\begin{aligned}
&\delta^*(f, g) = \frac{1}{16} \sum_{X,Y} |\rho^*_{XY}(f) - \rho^*_{XY}(g)|\\
&\text{where } \rho^*_{XY}(f) = \frac{f^*_{XY}}{f^*_Y f^*_Y} \text{and } f^*_X \text{ denotes the frequency of } X \text{ in } f.
\end{aligned} \tag{13}$$

The δ^* values ranges from 34 to 340. In general, lower δ^* values correspond to "closely related species" and higher values correspond to "distant species". We use F-measure to calculate the clustering accuracy. F-measure is a combination of precision and recall. Precision represents the fraction of reads within the cluster that belong to the same species. And recall is the extent to which the reads within a cluster belong to the same species. In order to obtain global performance statistics, we combined the F-measure by weighting each cluster by the number of reads in the cluster.

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[22]. Maximum likelihood favors more complex models leading to over-fitting and hence is unable to address this issue. Previously, 16s/18s rDNA have been used for phylotyping and assessing species diversity using a rare-fraction curve. Most methods rely on heuristics to guide the choices of clusters. Determining species diversity is still an active area of research and we do not address it in this paper.

The Bayesian method overcomes the bottleneck of Naive Bayes methods in their assumption of independence between the words in a genome. In Figure 2, we compare the performance of our Bayesian methods with its Naive Bayes counterparts over 450 datasets with δ^* values ranging from 34 to 340. We observe a positive correlation between δ^* and the accuracy of our methods, as also noted in [13]. The Bayesian method outperforms the Naive Bayes in all instances. A Bayesian model regards the word counts as being multinomially distributed and hence captures the correlation between words counts. However, Naive Bayes methods was found to be on an average twice as faster than the corresponding Bayesian methods. Increased accuracy on datasets with short reads in yet another consequence of Bayesian networks (Figure 2). Though the classification accuracy is correlated to the read length, the drop in accuracy (bounded by 5%) with the decrease in read length from 1000 bps to 50 bps is hardly significant.

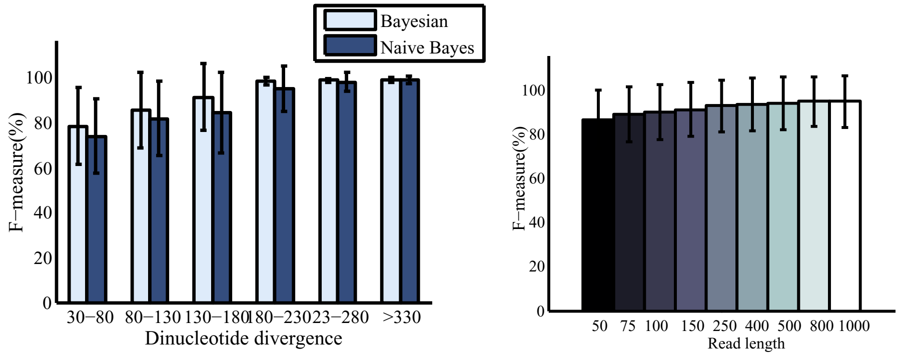

Fig. 2. Bayesian methods account for word overlaps. a) Comparison of performance of Bayesian mixture of Poissons and Multinomials with their Naive Bayes counterparts. We used a word length of 4. b) Effect of read length on clustering accuracy.

We systematically evaluated the robustness of our method to changes in the coverage ratio between species representative of various intergenomic differences. Binning results for 20 sets of simulated metagenomes with two species each is summarized in Figure 3.We varied the coverage ratio from 10:1 to 1:10 in stages, for the two species. We note that the accuracy drops at extreme coverages, when the fractional content of the species reduces to less than 10%.

Next, we analyzed the applicability of Bayesian mixture of Poissons in binning reads from low complexity communities, containing 2-6 species (Figure 3). The results were averaged over 50 datasets of varying divergences. Given that the multi-species dataset may contain reads from species with little intergenomic differences, there was a slight degradation in performance with the increase in number of species. This is in agreement with the results on variation with coverage, since the total coverage of each species is much lower in a multi-species dataset.

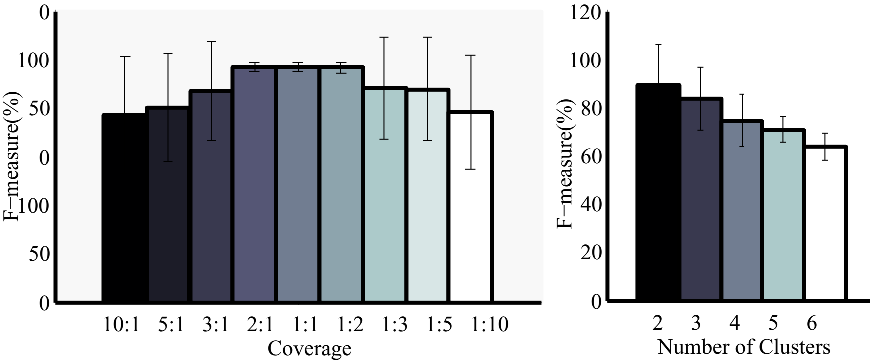

Fig. 3. Performance of Bayesian Poisson mixture model with varying coverage and different number of species (Word length of 4)

The above two observations are true of most composition based methods. Thus, even though composition based methods are suitable for binning reads from novel species, these methods by themselves are not sufficient for metagenomes containing a large number of species. For these methods to be practical on real metagenomes, we need to combine them with other similarity-based and abundance-based methods. In [12], the authors integrate the proposed composition based Scimm with similarity based methods to demonstrate increased performance on a medium-complexity simulated dataset. In our previous paper[20], we combined the Naive Bayes method with an abundance based method to characterize the Acid Mine Drainage dataset [23]. These results indicate that our method is suitable for binning reads belonging to dominant species, and that binning relatively rare species in a multi-species dataset may require modifications to the present Bayesian formulation.

In general, the discriminative power of the models increases with the length of words, despite the increasing space complexity. In our experiments, when we increased the word length from 2 to 7, initially the accuracy increases with word length. For word lengths beyond five, the accuracy begins to drop. This is because the feature matrix becomes high-dimensional and sparse. Hence, the model fails to predict the true feature distribution of different components. This necessitates dimension reduction before estimating the components in the model.

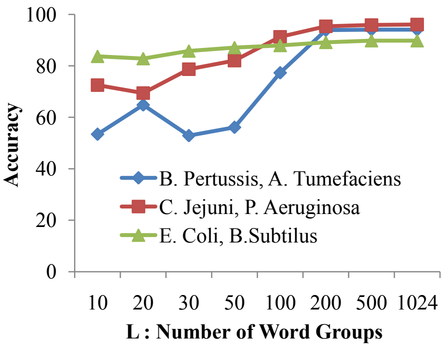

Fig. 4. Performance of Two-way Bayesian Poisson mixture model for values of word groups, L, varying from 10 to 1024. A word length of 5 is used.

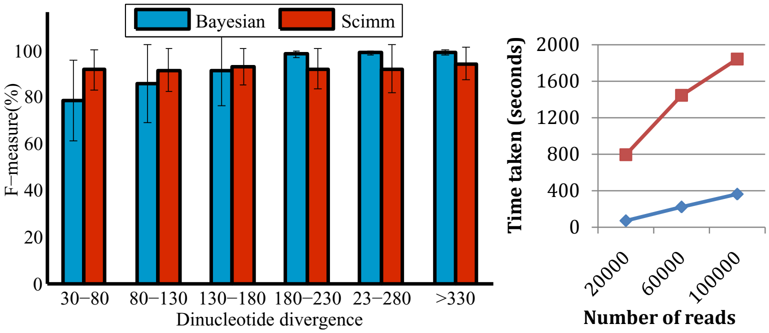

Fig. 5. Comparison of accuracy and time taken by Bayesian mixture of Poissons with Scimm. δ^* vary from 60 to 300. Read length of 200 bps and word length of 4.

In this paper, we perform "word grouping" to handle the above challenge. We propose a two-way mixture model where the mixture clusters induce a partition of the reads as well as of words. We used word length of 5 and varied the number of word groups from 10 to 1024 in stages (Figure 4). Performance stabilizes close to its optimal value at $L = 100$. This implies that the data can be classified using no more than ML dimensions, a significant reduction from the original number of dimensions. That is, the characteristic vectors are of a much lower dimension. 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.

Finally, in Figure 5, we compare the accuracy of our proposed multi-dimensional Bayesian model with state-of-art unsupervised composition-based method Scimm[12] averaged over 450 datasets. We used a read length of 100 bps and word length of 3. As the number of dimensions is relatively small, our method performs well without word grouping too. For δ^* values of 180 and above, our method performs better than Scimm. Though for δ^* values below 90, Scimm does better than our method. However, our method converges at least 5 times faster than Scimm. The reduced time taken by our method to achieve comparable results justifies its use for clustering large metagenome datasets.

5 Conclusion

In this paper, we proposed a multivariate Bayesian methods based on Poisson and Multinomial mixture model to cluster the reads in a metagenome by their species of origin. This work demonstrates the use of statistically based mixture models for analysis of metagenome datasets by suitable choices of probability distributions. The Poisson and Multinomial models can effectively cluster the reads when the word counts are very low. Bayesian networks are used to represent the conditional dependencies between the words. We examined the sensitivity of the method to the number of species, abundance profile and length of reads within the dataset. Much work needs to be done to validate the usefulness of these model for real metagenome datasets. Our method is an unsupervised method that does not require any training data. However, we still need to specify the number of species for the algorithm. A future direction for our work is to overcome this limitation. 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. The methods have been tested on metagenomes, but can be adapted for use with a variety of discrete sequence data such as document clustering data, web-logs, purchase history or stock market data among others.

References

1. Bailly-Bechet, M., Danchin, A., Iqbal, M., Marsili, M., Vergassola, M.: Codon Usage Domains over Bacterial Chromosomes. PLoS Comput. Biol. 2(4), e37+ (2006)
2. Bentley, S.D., Parkhill, J.: Comparative genomic structure of prokaryotes. Annual Review of Genetics 38(1), 771–791 (2004)
3. Brady, A., Salzberg, S.L.: Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nature Methods 6(9), 673–676 (2009)
4. Campbell, A., Mrázek, J., Karlin, S.: Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America 96(16), 9184–9189 (1999)
5. Chatterji, S., Yamazaki, I., Bai, Z., Eisen, J.: CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. ArXiv e-prints, 708 (August 2007)
6. Chen, K., Pachter, L.: Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS Comput. Biol. 1(2), e24 (2005)
7. Dalevi, D., Ivanova, N.N., Mavromatis, K., Hooper, S.D., Szeto, E., Hugenholtz, P., Kyrpides, N.C., Markowitz, V.M.: Annotation of metagenome short reads using proxygenes. Bioinformatics 24(16), i7–i13 (2008)

8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), 1–38 (1977)
9. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 1. Wiley (1968)
10. Heckerman, D.: A tutorial on learning with bayesian networks. Technical report, Learning in Graphical Models (1995)
11. Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C.: MEGAN analysis of metagenomic data. Genome research 17(3), 377–386 (2007)
12. Kelley, D., Salzberg, S.: Clustering metagenomic sequences with interpolated markov models. BMC Bioinformatics 11(1), 544 (2010)
13. Kislyuk, A., Bhatnagar, S., Dushoff, J., Weitz, J.S.: Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics 10(1), 316+ (2009)
14. Li, J., Zha, H.: Two-way poisson mixture models for simultaneous document classification and word clustering. Comput. Stat. Data Anal. 50, 163–180 (2006)
15. McHardy, A.C.C., Martín, H.G.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I.: Accurate phylogenetic classification of variable-length DNA fragments. Nature Methods 4(1), 63–72 (2007)
16. Rapp, M.S., Giovannoni, S.J.: The uncultured microbial majority. Annual Review of Microbiology 57(1), 369–394 (2003)
17. Reinert, G., Schbath, S., Waterman, M.S.: Probabilistic and Statistical Properties of Words: An Overview. Journal of Computational Biology 7(1-2), 1–46 (2000)
18. Robin, S., Rodolphe, F., Schbath, S.: DNA, Words and Models: Statistics of Exceptional Words. Cambridge University Press (2005)
19. Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., Sokhansanj, B.: Metagenome fragment classification using n-mer frequency profiles
20. Shruthi Prabhakara, R.A.: A two-way multi-dimensional mixture model for clustering metagenomic sequences. In: ACM BCB (2011)
21. Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., Glöckner, F.O.: Application of tetranucleotide frequencies for the assignment of genomic fragments. Environmental Microbiology 6(9), 938–947 (2004)
22. Tibshirani, R., Walther, G.: Cluster Validation by Prediction Strength. Journal of Computational & Graphical Statistics 14(3), 511–528 (2005)
23. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., Banfield, J.F.: Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978), 37–43 (2004)
24. Willse, A., Tyler, B.: Poisson and multinomial mixture models for multivariate sims image segmentation. Analytical Chemistry 74(24), 6314–6322 (2002)