Model based clustering analysis of 16S rRNA sequence

Wei Zhang[1], Rui Kuang[1], and Baolin Wu[2⋆]

[1]Department of Computer Science and Engineering
[2]Division of Biostatistics, School of Public Health
University of Minnesota Twin Cities

Abstract. With the advances in next-generation sequencing technology, 16S ribosomal RNA (rRNA) gene sequences are increasingly being used to interpret the phytogenetic relationship between organism. A common research question is the taxonomy independent analysis where sequences are clustered into operational taxonomic units (OTUs) to estimate microbial communities in given environmental samples. Most existing clustering methods for 16S rRNA gene sequences are computing intensive and the results often subjectively depend on some user-set parameters, which often leads to a wide range of clustering results that are hard to interpret. To address these challenges, we propose a flexible and efficient gaussian mixture model combined with multidimensional scaling (called GMM-MDS) to cluster 16S rRNA sequences for OTU prediction. The proposed model requires no user-set parameters and produces the clustering results objectively and adaptively based on the data. Through applications to public datasets and numerical simulations, we demonstrate that the proposed method performs competitively compared to the existing methods.

1 Introduction

In recent years, with the rapid development of 16S rRNA next-generation sequencing technology, researchers can easily sequence millions of signature sequences from microbial communities in an acceptable time with a reasonable cost, which provide new approaches to many metagenomic problems. A major goal of these studies is to uncover the true structure of microbial community in given environment samples. Many computational methods have been proposed to cluster sequences into operational taxonomic units (OTUs) to approximate the number of microbial communities which enable microbiologists to keep pace with the growth of huge amount of genomic data sets available today. However, most existing clustering methods depend on some user-set parameters, and typically a small change in the algorithm parameters can lead to significantly varying clustering results [1–3], which makes the results hard to interpret and often requires microbiologists to put more efforts to manually and subjectively choose the parameters.

Current taxonomy independent clustering algorithms can be grouped into several subcategories as introduced in [2, 4]. The three widely used approaches are hierarchical clustering algorithm (HC) (e.g., ESPRIT-Tree [5], DOTUR [6], mothur [7]), greedy heuristic clustering algorithm (GHC) (e.g., UCLUST [8], CD-HIT [9]), and Bayesian clustering algorithm (BC) (e.g., CROP [3], BEBaC [4]). Most HC algorithms construct a hierarchical tree based on a computed distance matrix between all sequences, then a user-specified cutoff is applied to assign the leaves in the tree to different OTUs. For example, ESPRIT-Tree [5], an efficient HC algorithm, represents each cluster of sequences as a probabilistic sequence to avoid extensive computation of pairwise distances between clusters, and reports the clustering results for varying distance cutoffs (e.g., from 0.01 to 0.15) in a single run. However, even for microbiologists with enough background knowledge, the 'optimal' cutoff is hard to be decided. Furthermore, the reported results by ESPRIT-Tree typically varies from different runs, which further complicates the interpretation. Similar drawbacks have been observed in CROP [3], a BC algorithm, which splits the sequences into random blocks and applies Bayesian clustering algorithms to each block. And the lower and upper bound distance parameters need to be specified before running the program. The CROP proposed the split and merge process to partially resolve the randomness of clustering, but different runs still lead to largely different results. GHC approaches incorporated

⋆ Email: baolin@umn.edu

a similar partition strategy, which however is not guaranteed to approximate the true structure of microbial communities [2].

In this paper we developed a flexible and efficient taxonomy independent clustering model to address some of the limitations of the existing methods. When analyzing two simulation datasets and five 16S rRNA gene sequence datasets, the results reported by our model are comparable to the best of the other methods across a range of user-set parameters.

2 Methods

In this section, we first describe the construction of distance matrix for 16S rRNA gene sequence data by using the pairwise alignment algorithm. We then apply the multidimensional scaling (MDS) based dimension reduction to the distance matrix to construct an Euclidean coordinate matrix. A gaussian mixture model is then fitted to the coordinate matrix. And the Bayesian Information Criterion (BIC) is used to adaptively and subjectively choose the dimensions of the coordinate matrix and the number of OTUs.

Assume that there are n read sequences $\boldsymbol{r} = (r_1, ..., r_n)$ in total in the 16S rRNA data.

2.1 Pairwise alignment and dissimilarity matrix calculation

The Needleman-Wunsch algorithm [10] was used for pairwise alignment and we followed the strategies described in CROP [3] and Quickdist algorithm [6] to calculate the dissimilarity matrix $\boldsymbol{D} = \{d_{ij}|i, j = 1, 2, ..., n\}$ for the sequence pairs. The dissimilarity d_{ij} is calculated as a percentage number of mismatches in the pairwise alignment between r_i and r_j.

2.2 Multidimensional scaling (MDS) and efficient parallel computation

We first apply the classical MDS procedure [11] to the dissimilarity matrix $\boldsymbol{D}$ to project the data into a $n \times p$ dimensional Euclidean coordinate matrix $\boldsymbol{X}$. Here p is typically much smaller than n. Intuitively the first several dimensions of $\boldsymbol{X}$ will capture most of the information in the dissimilarity matrix $\boldsymbol{D}$. Our intuitive idea is to analyze the first several dimensions of $\boldsymbol{X}$ using a model-based clustering algorithm, and then efficiently and adaptively choose the dimension and the number of clusters (i.e., OTUs) automatically based on the data using a model selection criterion. Note that, $\boldsymbol{D}$ is not a Euclidean distance matrix, and some eigenvalues can be negative. We have observed that the negative eigenvalues are small in magnitude (Figure 1A), and generally only the first several positive eigenvalues are large (Figure 1B). Thus we can approximate the dissimilarity matrix reasonably well using the first several dimensions of $\mathbf{X}$.

For an extremely large dataset with hundreds of thousands sequences, parallel computing system, such as Hadoop [12] or MPI [13], can be applied to parallelize an elegant algorithm, SMACOF (Scaling by MAjorizing a COmplicated Function) [14], to efficiently compute MDS solution. A previous method proposed in [15] analyzed the performance of block matrix decomposition for parallel SMACOF implementation on multicore cluster system. Since only first several dimensions ($\sim$10) of $\boldsymbol{X}$ are informative, parallel SMACOF can efficiently compute and export the result of MDS.

2.3 Gaussian mixture model for clustering

We apply a gaussian mixture model (GMM, [16]) to the 16S rRNA sequence data. Specifically, for the first k dimensions of matrix $\boldsymbol{X}$, denoted as $\boldsymbol{Y}$, where $\boldsymbol{Y}_i = (y_{i1}, ..., y_{ik})$, and given the number of components G, we propose the following gaussian mixture model

$$\Pr(\boldsymbol{Y}_i|\boldsymbol{\theta}, \boldsymbol{\mu}, \sigma^2\boldsymbol{I}) = \sum_{g=1}^{G} \theta_g f(\boldsymbol{Y}_i|\boldsymbol{\mu}_g, \sigma^2\boldsymbol{I}), \quad \sum_{g=1}^{G} \theta_g = 1, \quad \theta_g \geq 0, \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_G)$ are the mixing proportions. The component means $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G)$ are cluster and dimension specific, where $\boldsymbol{\mu}_g = (\mu_{g1}, ..., \mu_{gk})$. Here we have assumed an equal-volume spherical covariance matrix $\sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is an $k \times k$ identity matrix. The individual component multivariate normal density is

$$f(\boldsymbol{Y}_i|\boldsymbol{\mu}_g, \sigma^2\boldsymbol{I}) = (2\pi)^{-\frac{k}{2}}\sigma^{-k}\exp\Big\{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}(y_{ij}-\mu_{gj})^2\Big\} \tag{2}$$

We estimate the model parameters by maximizing the sequence data log likelihood

$$\mathcal{L} = \sum_{i=1}^{n}\log\Big\{\sum_{g=1}^{G}\theta_g f(\boldsymbol{Y}_i|\boldsymbol{\mu}_g, \sigma^2\boldsymbol{I})\Big\}. \tag{3}$$

Standard EM algorithm can be applied to obtain the maximum likelihood estimator (MLE) [16]. The gaussian mixture model typically has many local maxima and multiple random initial values are applied to find a good solution. Here we propose a parameter initialization based on the k-means++ algorithm [17].

The component centers are initialized following the setup in k-means++ as follows. The first component center is chosen uniformly at random from the sequences that are being clustered, after which each subsequent component center is chosen from the remaining sequences with probability proportional to its squared distance from the sequence's closest existing component center. In our numerical studies, this parameter initialization method yields considerable improvement in finding the MLE for the gaussian mixture model.

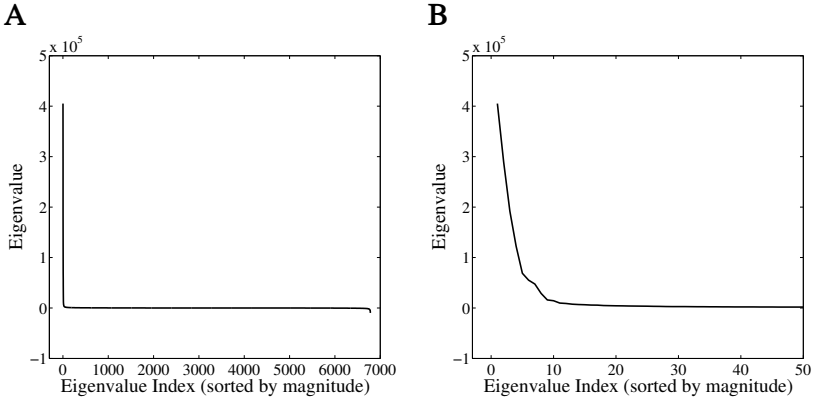
2.4 Model selection via BIC

We apply the Bayesian Information Criterion (BIC) [18] to select the matrix dimension k and component number G (hence OTUs). For the proposed mixture model, the BIC is computed as a penalized maximum log likelihood accounting for the number of parameters (model complexity),

$$BIC = 2\mathcal{L}^* - (k+1)G\log(n), \tag{4}$$

where $\mathcal{L}^*$ is the maximized log likelihood in equation (3). Then we pick up dimension k and component number G based on the largest BIC. Given the estimated normal mixture model, the posterior probability, $\tau_{ig} = \theta_g f(\boldsymbol{Y}_i|\boldsymbol{\mu}_g, \sigma^2\boldsymbol{I})/\Pr(\boldsymbol{Y}_i)$, probabilistically quantifies the relative chance of sequence belonging to each OTU, and is used to assign each sequence to the OTU with the maximum posterior probability, $\arg\max_g \tau_{ig}$.

Fig. 1. Eigenvalues of $\boldsymbol{D}$ for the eMC dataset. (A) All the eigenvalues of dissimilarity matrix $\boldsymbol{D}$ sorted by magnitude from largest to smallest (6782 total). (B) The first 50 largest eigenvalues of $\boldsymbol{D}$.



3 Results

In the experiments, we compare the proposed GMM-MDS to the following three representative OTU prediction methods, CROP, ESPRIT-Tree and UCLUST.

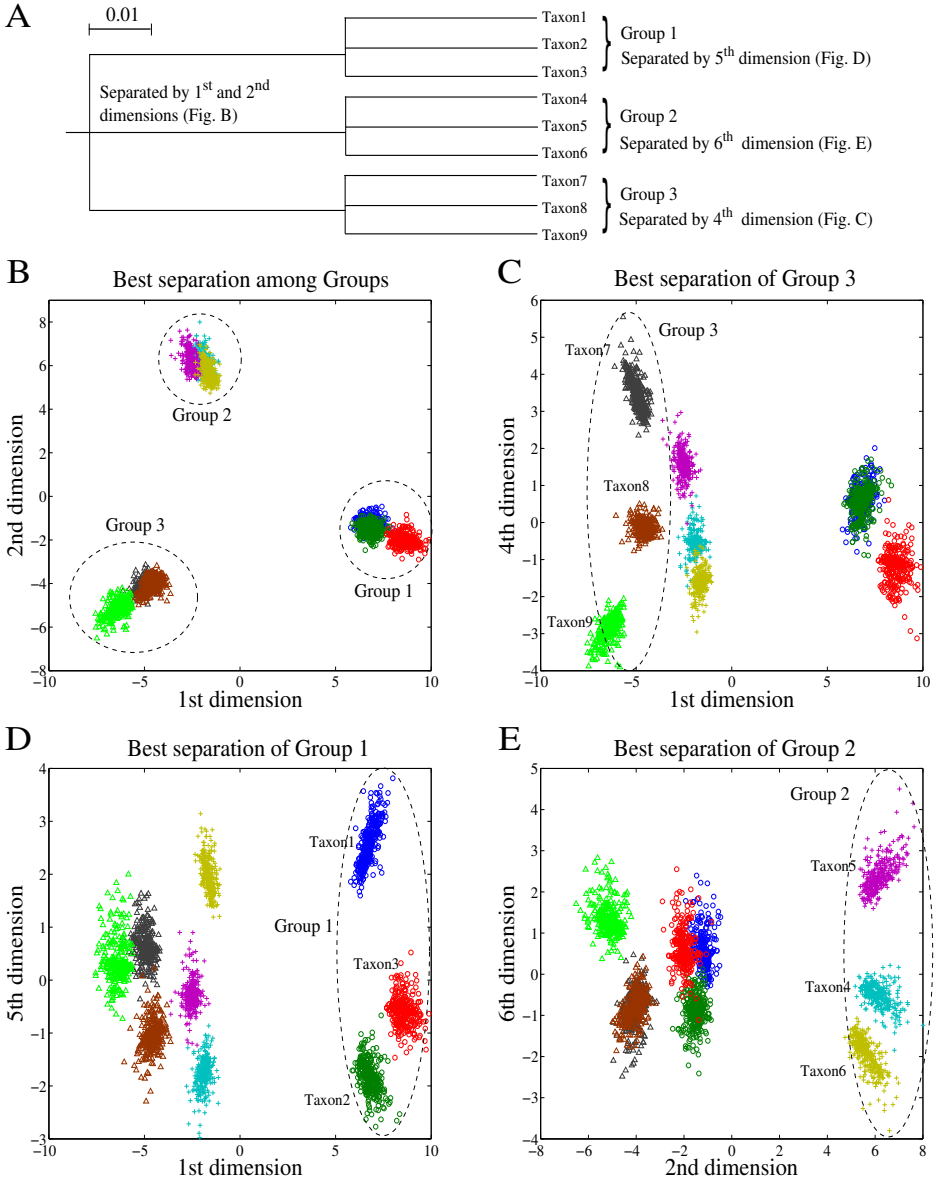
3.1 Simulation study

Follow the steps described in [4], we simulated two datasets. The first one contains 9 taxonomic unites that were designed as shown in Figure 2(A). The designed phytogenetic tree with a simple structure contains two distances levels, 6% and 14%. Then, 9 consensus sequences each with 500 base pairs were generated using the software Seq-Gen [19] according to the phytogenetic tree. We simulate 500 individual sequences from each consensus sequence, the average deviation of a sequence from the consensus sequence was δ. Specifically, a random number is drawn from a gaussian distribution $N(0, \delta)$, the absolute value of the number is the percentage of mismatches between an individual sequence and the consensus sequence. The δ was set as 3% for all 9 taxonomic unites in this dataset. In total we simulated 4500 sequences each with 500 base pairs generated from the 9 taxonomic unites. Similarly, we simulated the second dataset consisting of 12 taxonomic unites that were designed as shown in Figure 3. The designed phytogenetic tree was composed of a more complex structure, which contains multiple distances levels with 1% minimum distance and 30% maximum distance. δ_i (i=1,2,...,12), which are set as {2,2,3,3,4,5,3,2,3,4,3,3}% for the 12 taxonomic unites. We generated 500 sequences from each consensus sequences (500 bp), and 6000 sequences in total. We ran our model on each dataset with dimension k from 2 to 10 and component G from 2 to 100. For each specific number of components G, we ran the model 100 times with different initials selected based on the k-means++ and report the largest BIC, then pick up the k and G based on the largest BIC for each dataset.

The results of clustering the 4500 and 6000 simulated sequences by different models are reported in Table 1 and Table 2, respectively. Normalized mutual information (NMI) score [20] and adjusted Rand index (ARI) [21] are used to evaluate how the clustering outcomes from each model agree with the ground truth. NMI/ARI=1 means the clustering result is the same as the ground truth, and NMI/ARI=0 means the sequences are randomly grouped. From Table 1 and Table 2 we can see that the clustering results vary significantly across different parameters for the UCLUST, ESPRIT-Tree and CROP. For CROP and ESPRIT-Tree, when the cut-off is low, it tends to over-estimate the number of OTUs and report many false positive ones; when the cut-off is high, it tends to under-estimates the number of OTUs and assign the closely related taxonomic units into one OTU. The GHC algorithm UCLUST does not work well under a wide range of parameter settings. In general the CROP performs better than UCLUST and ESPRIT-Tree. When considering estimates of both the number of OTUs and clustering accuracy based on NMI/ARI, the proposed GMM-MDS is comparable to the best results of CROP.

To assess the effectiveness of MDS for capturing the dissimilarity information, we evaluate whether the first several dimensions of the projected matrix capture the major structure of the phytogenetic tree in Figure 2(A). In Figure 2(B)-(E), we show the scatter plots of combinations of the first several dimensions after applying MDS to the dissimilarity matrix. It is clear that the first two dimensions capture the information to separate the three groups (Figure 2(B)) and 4th, 5th and 6th dimensions contain the information to distinguish the taxonomic unites within group 3, 1 and 2, respectively. From the first several dimensions estimated by MDS, we can easily reconstruct the phytogenetic tree. The similar analysis on the second simulated data is shown in Figure 7 in **Appendix**. For the more complex real datasets, we therefore expect that the MDS can provide valuable visualization information regarding the underlying structure of the microbial communities.

Fig. 2. Relations between phytogenetic tree and MDS. (A) Phytogenetic tree of 9 taxonomic unites within 3 groups. (B)-(E) Scatter plots of first several dimensions with largest eigenvalues on the first simulated dataset. Taxonomic unites and groups are distinguished by colors and markers, respectively.



3.2 16S rRNA gene sequence datasets preparation

Five 16S rRNA gene sequence datasets were used in the experiment to numerically compare different methods for their estimation of the number of OTUs. Datasets Clone43A and Clone43B were described by [22] and preprocessed following [3]. They are generated from 43 16S rRNA templates with at least 3% difference from each other. Clone43A contained only the reads that were within 3% of one of the 43 templates and Clone43B contained all the reads. Human skin microbiome dataset Grice was generated by [23] and preprocessed following [3]. The Grice data contained 33 genera classified by the ribosomal database project classifier [24], which were considered as the ground truth. The Quince dataset [25] contained 10852 unique reads sequenced from V5 and V6 regions of 90 different clones of bacteria. The even composition Mock Communities (eMC) dataset was generated by [26] and preprocessed following [4]. We extracted 10% of the reads from the data which consists of 6782 unique reads sequenced from

Fig. 3. Phytogenetic tree of 12 taxonomic units.

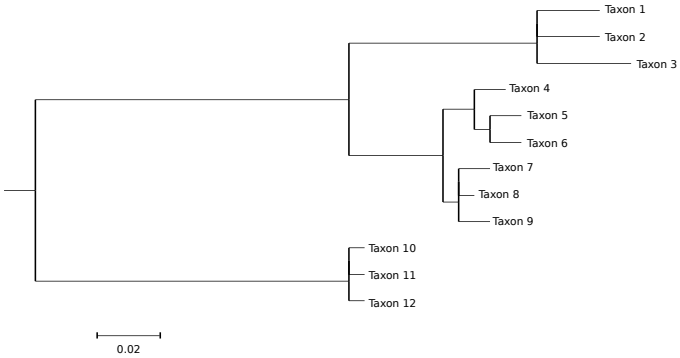


Table 1. Results for the simulated dataset with 9 taxonomic units.

Algorithm (parameters)	estimated number of OTUs	NMI	ARI
GMM-MDS	29	0.8334	0.7681
UCLUST(1%)	3865	0.4273	0.0064
UCLUST(3%)	2604	0.4786	0.0748
UCLUST(5%)	1356	0.5742	0.2959
UCLUST(10%)	173	0.7876	0.6763
ESPRIT-Tree(0.03)	2294	0.5512	0.3459
ESPRIT-Tree(0.05)	1126	0.7099	0.6894
ESPRIT-Tree(0.08)	255	0.8467	0.7511
ESPRIT-Tree(0.10)	198	0.6266	0.3887
ESPRIT-Tree(0.12)	69	0.6521	0.3960
ESPRIT-Tree(0.15)	14	0.4480	0.1808
CROP(1%)	1106	0.6580	0.4126
CROP(2%)	321	0.8084	0.6820
CROP(3%)	443	0.8381	0.8316
CROP(4%)	87	0.9530	0.9561
CROP(5%)	10	0.7997	0.6194
CROP(6%)	4	0.6410	0.3897
Ground truth	9		

Table 2. Results for the simulated datasets with 12 taxonomic unites.

Algorithm (parameters)	estimated number of OTUs	NMI	ARI
GMM-MDS	40	0.6264	0.4085
UCLUST(1%)	5287	0.4551	0.0047
UCLUST(3%)	3519	0.5022	0.0942
UCLUST(5%)	2091	0.5464	0.3336
UCLUST(10%)	448	0.6215	0.3057
ESPRIT-Tree(0.03)	3182	0.5340	0.2833
ESPRIT-Tree(0.05)	1873	0.6255	0.4721
ESPRIT-Tree(0.08)	924	0.5228	0.2483
ESPRIT-Tree(0.10)	521	0.5472	0.2566
ESPRIT-Tree(0.12)	186	0.5721	0.2597
ESPRIT-Tree(0.15)	88	0.3701	0.1019
CROP(1%)	1693	0.6282	0.3568
CROP(2%)	982	0.6678	0.4930
CROP(3%)	930	0.6341	0.3948
CROP(4%)	358	0.6043	0.3062
CROP(5%)	20	0.5877	0.2628
CROP(6%)	4	0.5774	0.2581
Ground truth	12		

22 bacterial species and 28 reference sequences. Table 3 summarizes the five dataset with their known number of OTUs.

We ran our model on each dataset with dimension k from 2 to 10 and component G from 2 to 150 for dataset Quince and 2 to 100 for the other ones. For each specific number of components G, we ran the model 100 times and report the largest BIC, then pick up k and G based on the largest BIC for each dataset. In Table 4, we report the optimal number of OTUs with the

Table 3. Datasets.

Datasets	Clone43A	Clone43B	Grice	Quince	eMC
number of unique reads	5296	9443	1009	10852	6782
known number of OTUs	43	43	33	90	28

corresponding BIC for each dimension k, the best one across all the dimensions are bolded. From the results we can see that for four out of five datasets, we get the optimal number of OTUs when dimension $k = 2$, which means that the first two dimensions of Euclidean coordinates capture most of the dissimilarity information for estimating the structure of microbial communities. The scatter plots in Figure 4 show the first two dimensions with the estimated cluster centers of the four datasets. Besides that, we plot the eMC dataset with its 'true label', 22 bacterial species, with the first two and first three dimensions with the largest eigenvalues in Figure 5(A) and (B), respectively.

In Figure 6, we plot the BIC across different number of components G for the optimal dimension k. The red marker is the estimated G. Note that, we only consider those components with more than 4 sequences as an OTU. In most cases, the reported number of components G in the plots are larger than the optimal number of OTUs.

Table 4. Number of OTUs estimated by GMM-MDS with dimensions k from 2 to 10.

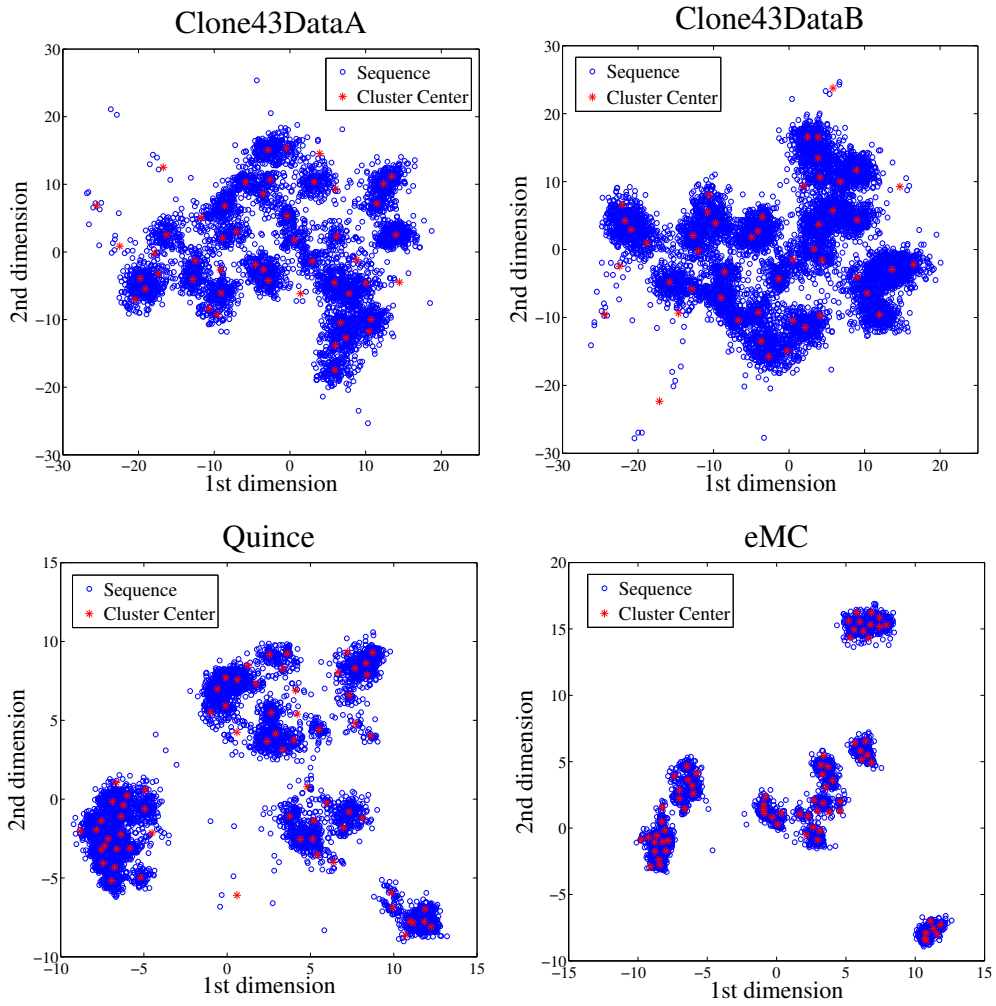
Dataset	**Clone43DataA**		**Clone43DataB**		**Grice**		**Quince**		**eMC**	
Dimension(k)	# of OTUs	BIC	# of OTUs	BIC	# of OTUs	BIC	# of OTUs	BIC	# of OTUs	BIC
2	**49**	**-67942.1**	**48**	**-121118.1**	27	-2712.0	**79**	**-83648.1**	**77**	**-45185.9**
3	85	-90554.9	85	-164109.5	34	-2679.4	130	-103165.4	81	-46978.8
4	78	-112260.3	87	-203241.4	39	-1996.4	134	-120679.2	87	-49363.2
5	87	-132823.6	90	-241242.8	43	-1589.7	128	-135761.5	89	-53911.3
6	83	-152355.2	89	-277861.0	38	-1608.3	130	-154181.3	84	-60294.9
7	77	-173297.2	93	-315085.9	36	-1100.3	134	-171949.7	87	-65411.7
8	86	-193005.1	85	-352972.6	41	-871.2	131	-188590.2	93	-70180.0
9	85	-212906.2	91	-390824.7	**38**	**-209.1**	128	-204811.6	95	-82259.6
10	85	-232972.5	93	-427240.4	39	-364.6	129	-220762.3	94	-93780.5

Table 5 summarizes the results of OTU estimation by the other methods compared to GMM-MDS. In most of the datasets, the numbers of OTUs estimated by GMM-MDS are very close to the ground truth.

Table 5. Number of OTUs estimated by all the methods.

	Datasets				
Algorithms(parameters)	Clone43A	Clone43B	Grice	Quince	eMC
GMM-MDS	49	48	38	79	77
UCLUST(1%)	3288	6002	149	2372	730
UCLUST(3%)	246	1763	59	133	79
UCLUST(5%)	89	260	40	79	36
UCLUST(10%)	46	79	29	42	19
ESPRIT-Tree(0.03)	533	1880	52	89	34
ESPRIT-Tree(0.05)	46	242	40	54	23
ESPRIT-Tree(0.08)	42	55	31	37	16
ESPRIT-Tree(0.10)	42	45	26	26	16
ESPRIT-Tree(0.12)	39	42	20	19	14
ESPRIT-Tree(0.15)	38	39	15	12	10
CROP(1%)	286	1373	84	193	106
CROP(2%)	61	334	58	63	31
CROP(3%)	50	109	42	43	22
CROP(4%)	44	71	37	30	21
CROP(5%)	44	53	21	19	15
CROP(6%)	43	44	16	13	14
Ground truth	43	43	33	90	28

Fig. 4. Estimated cluster centers by GMM-MDS.



4 Discussion

The proposed GMM-MDS for OTU prediction is motivated by the novel CROP method [3], which is based on the birth-death process Bayesian mixture modeling. GMM-MDS had several improvements compared to the CROP. (1) We utilize all the pairwise sequence distances and do not need partition into random blocks, and hence can truly capture the underlying OTU. (2) By projecting the dissimilarity matrix into Euclidean coordinates to be fitted by the mixture model, we can construct a truly likelihood based approach, which enables a principled approach to predicting the number of OTUs efficiently based on the well-established model selection criterion. (3) We adapt the recently proposed remarkable k-means++ approach [17] to solve the gaussian mixture model. In general finding the k-means solution is NP-hard and most existing methods are heuristic leading to local maximum, while the k-means++ can reach the global maximum with theoretical guarantee. In our limited numerical studies, we have found that initializing the gaussian mixture model based on the k-means++ approach often leads to solutions with the maximum log likelihood. (4) When using a large number of dimensions in the covariate matrix $\boldsymbol{X}$, we could exactly reproduce the dissimilarity matrix $\boldsymbol{D}$ (property of the MDS algorithm), and the proposed GMM-MDS can then be roughly treated as an improved frequentist version of CROP. Intuitively the computed pairwise sequence distance is discrete, noisy, incomplete, and scattered. By selecting appropriate number of dimensions, we could best capture the dissimilarity information that truly contributes to the OTU differences. In

Fig. 5. Scatter plots of eMC dataset. The 22 bacterial species are labeled with different colors and markers. (A) Scatter plot of the first two dimensions with the largest eigenvalues. (B) Scatter plot of the first three dimensions with the largest eigenvalues.

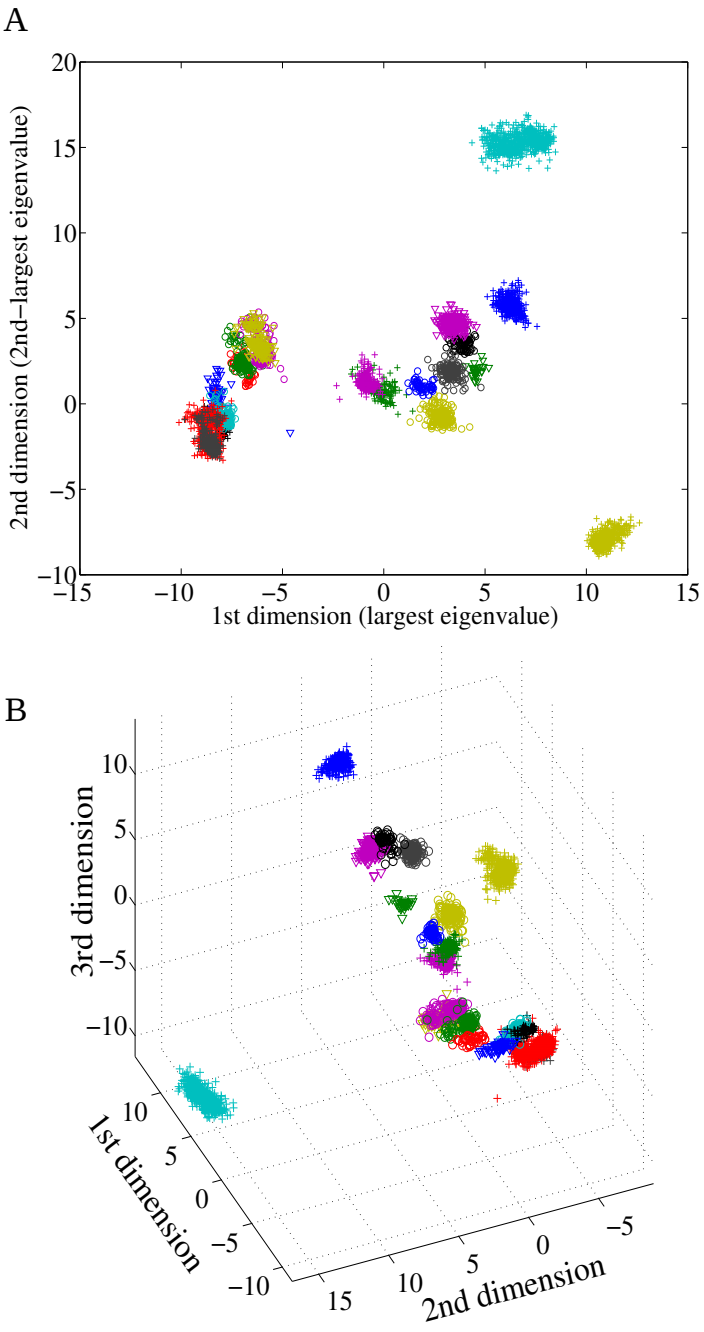
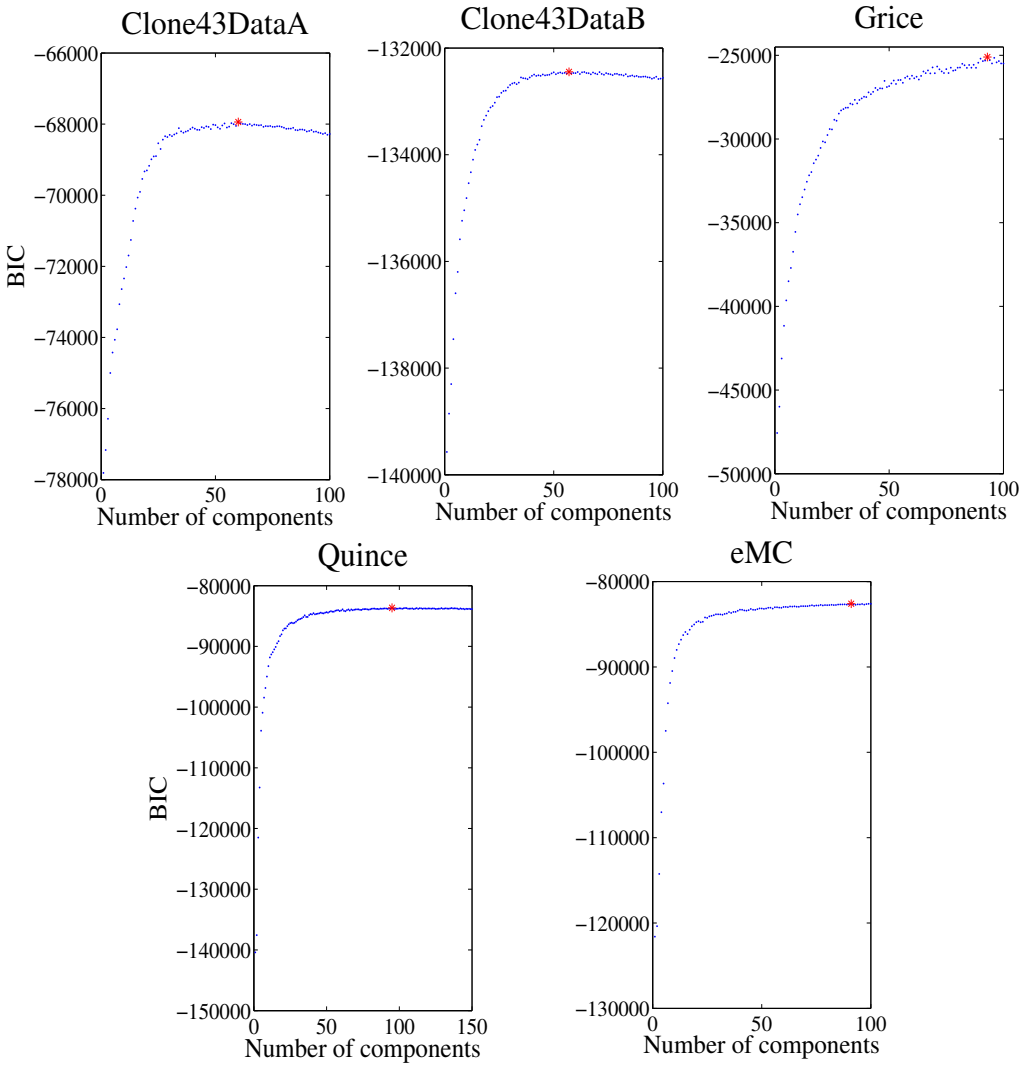


Fig. 6. Estimated BIC for different dataset. The red marker is the estimated number of components.



addition, by applying parallel/cloud computing to the MDS step to avoid directly handling the full dissimilarity matrix, GMM-MDS is, in principle, scalable to huge datasets in practical metagenomics problems.

The proposed approach of projecting dissimilarity metric into Euclidean coordinates to be incorporated into further statistical model building is a novel idea that transcends the metagenomic OTU prediction problem [27]. Similar approach has been successfully applied to the study of disease risk and protein function inferences [28–30].

In the current paper, we have directly used the MDS of the calculated sequence dissimilarity matrix $\boldsymbol{D}$ for Euclidean coordinates calculation. Note that $\boldsymbol{D}$ is not guaranteed to be a distance matrix (i.e., non-negative definite). Nonetheless the current approach has performed consistently well in our numerical studies. A further research topic is to build a regularized kernel matrix in an RKHS based on the approach of [30], which could enable us to build a kernel based nonlinear clustering approach. Some preliminary results are promising and we will report the results elsewhere in the future.

Acknowledgments. This research was supported in part by NIH grant GM083345.

References

1. White, J., Navlakha, S., Nagarajan, N., Ghodsi, M.-R., Kingsford, C., and Pop, M. (2010) Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. *BMC Bioinformatics,* **11**(1), 152.
2. Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X., and Mai, V. (2012) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in Bioinformatics,* **13**(1), 107–121.
3. Hao, X., Jiang, R., and Chen, T. (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics,* **27**(5), 611–618.
4. Cheng, L., Walker, A. W., and Corander, J. (2012) Bayesian estimation of bacterial community composition from 454 sequencing data. *Nucleic Acids Research,* **40**(12).
5. Cai, Y. and Sun, Y. (2011) ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Research,* **39**(14), e95.
6. Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006) Microbial diversity in the deep sea and the underexplored rare biosphere. *Proceedings of the National Academy of Sciences,* **103**(32), 12115–12120.
7. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology,* **75**(23), 7537–7541.
8. Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* **26**(19), 2460–2461.
9. Li, W., Jaroszewski, L., and Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics,* **17**(3), 282–283.
10. Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology,* **48**(3), 443–453.
11. Kruskal, J. and Wish, M. (1978) Multidimensional Scaling, SAGE Publications, .
12. White, T. (2009) Hadoop: The Definitive Guide, O'Reilly, first edition edition.
13. Gabriel, E., Fagg, G. E., Bosilca, G., Angskun, T., Dongarra, J. J., Squyres, J. M., Sahay, V., Kambadur, P., Barrett, B., Lumsdaine, A., Castain, R. H., Daniel, D. J., Graham, R. L., and Woodall, T. S. (2004) Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation. *Proceedings, 11th European PVM/MPI Users' Group Meeting,* pp. 97–104.
14. Borg, I. and Groenen, P. (2005) Modern Multidimensional Scaling: Theory and Applications, Springer, .
15. Parallel Data Mining from Multicore to Cloudy Grids Cetraro, Italy (2008).
16. Fraley, C. and Raftery, A. E. (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association,* **97**(458), 611–631.
17. Arthur, D. and Vassilvitskii, S. (2007) k-means++: the advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms,* pp. 1027–1035.
18. Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics,* **6**, 461–464.
19. Rambaut, A. and Grass, N. C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences : CABIOS,* **13**(3), 235–238.
20. Studholme, C., Hill, D. L. G., and Hawkes, D. J. (1999) An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition,* **32**(1), 71–86.
21. Hubert, L. and Arabie, P. (1985) Comparing partitions. *Journal of classification,* **2**(1), 193–218.
22. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol,* **8**(7), R143.
23. Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., NISC Comparative Sequencing Program, Bouffard, G. G., Blakesley, R. W., Murray, P. R., Green, E. D., Turner, M. L., and Segre, J. A. (May, 2009) Topographical and temporal diversity of the human skin microbiome. *Science,* **324**(5931), 1190–1192.
24. Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T. L., Garrity, G. M., and Tiedje, J. M. (2009) The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research,* **37**, 141–145.
25. Quince, C., Lanzen, A., Curtis, T., Davenport, R., Hall, N., Head, I., Read, L., and Sloan, W. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods,* **6**, 639–641.
26. Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., Methé, B., DeSantis, T. Z., Consortium, T. H. M., Petrosino, J. F., Knight, R., and Birren, B. W. (March, 2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research,* **21**(3), 494–504.
27. Wahba, G. (2010) Encoding Dissimilarity Data for Statistical Model Building. *Journal of statistical planning and inference,* **140**(12), 3580–3596 PMID: 20814436.

28. Hou, J., Jun, S.-R., Zhang, C., and Kim, S.-H. (2005) Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National Academy of Sciences of the United States of America,* **102**(10), 3651–3656 PMID: 15705717.
29. Bravo, H. C., Lee, K. E., Klein, B. E. K., Klein, R., Iyengar, S. K., and Wahba, G. (2009) Examining the relative influence of familial, genetic, and environmental covariate information in flexible risk models. *Proceedings of the National Academy of Sciences,* **106**(20), 8128–8133 PMID: 19420224.
30. Kong, J., Klein, B. E. K., Klein, R., Lee, K. E., and Wahba, G. (2012) Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases, and mortality. *Proceedings of the National Academy of Sciences,* **109**(50), 20352–20357 PMID: 23175793.

Appendix

Fig. 7. Relations between phytogenetic tree and MDS. (A) Phytogenetic tree of 12 taxonomic unites within 3 groups. (B)-(H) Scatter plots of first several dimensions with largest eigenvalues on the seccond simulated dataset. Taxonomic unites and groups are distinguished by colors and markers, respectively.