Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 6598307, 11 pages
http://dx.doi.org/10.1155/2016/6598307

Research Article

Motif-Based Text Mining of Microbial Metagenome Redundancy Profiling Data for Disease Classification

Yin Wang,[1,2] Rudong Li,[3] Yuhua Zhou,[4] Zongxin Ling,[5] Xiaokui Guo,[4] Lu Xie,[2] and Lei Liu[1,2]

[1]*Shanghai Public Health Clinical Center and Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China*
[2]*Shanghai Center for Bioinformation Technology, Shanghai 201203, China*
[3]*Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[4]*Department of Medical Microbiology and Parasitology, Institutes of Medical Sciences, Shanghai Jiao Tong University School of Medicine, Shanghai 200240, China*
[5]*Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310003, China*

Correspondence should be addressed to Lei Liu; liulei@fudan.edu.cn

Received 28 October 2015; Accepted 12 January 2016

Academic Editor: Zhenguo Zhang

Background. Text data of 16S rRNA are informative for classifications of microbiota-associated diseases. However, the raw text data need to be systematically processed so that features for classification can be defined/extracted; moreover, the high-dimension feature spaces generated by the text data also pose an additional difficulty. *Results*. Here we present a Phylogenetic Tree-Based Motif Finding algorithm (PMF) to analyze 16S rRNA text data. By integrating phylogenetic rules and other statistical indexes for classification, we can effectively reduce the dimension of the large feature spaces generated by the text datasets. Using the retrieved motifs in combination with common classification methods, we can discriminate different samples of both pneumonia and dental caries better than other existing methods. *Conclusions*. We extend the phylogenetic approaches to perform supervised learning on microbiota text data to discriminate the pathological states for pneumonia and dental caries. The results have shown that PMF may enhance the efficiency and reliability in analyzing high-dimension text data.

1. Introduction

The microbial ecology in human determines or promotes necessary bioprocesses in human bodies, and compositions of microbial communities can be reflections for the health conditions of the hosts [1]. In fact, the complex microbial communities play key roles in human health from time to time. For example, dysfunction of microbiota biogeography or infection of pathogenic microbiota would lead to a series of human diseases, like pneumonia [2], dentes cariosus [3], and so on [4, 5]. Fortunately, sequencing of 16S rRNA provides informative knowledge for the distributions of microbiota [6]. For instance, microbiota taxonomy analysis based on the sequence data by bioinformatic tools such as Ribosomal Database Project (RDP) website would facilitate investigations of key microorganisms associated with certain host diseases [7].

On the other hand, traditional (supervised) methods such as feature selection are frequently adopted in classifications of microbiota-associated disease samples, for example, selecting the microorganism(s) which can maximally discriminate diseased and healthy hosts [6]. Nonetheless, substantial amounts of sequence data actually embody the characteristics of entire microbial communities rather than individual microbes [8]. Hence, the mapping results of 16S rRNA segments to individual microbes based on the sequencing data would not be informative enough for the aftermath feature selection. Furthermore, sequences that cannot be mapped to known

microbes might also have certain importance. Therefore, algorithms focusing on the textual features of microbiota sequences themselves (e.g., k-mer/k-tun features) have been applauded by recent researchers, as they skip the sequence-microbe mapping and hence avoid the intrinsic drawbacks [9, 10].

However, abundantly many features can be defined regarding raw text data (i.e., strings); in other words, the dimension of feature space would usually be extremely high; thus the "curse of dimensionality" resulted [11]. In this regard, motif-oriented algorithms are capable of accelerating the feature selection pipeline, as generalizing or lumping the textual features of a lot of strings into certain motifs is equivalent to degenerating the feature space (i.e., dimension reduction) [12, 13]. Nonetheless, extracting the motifs put forward another issue, intuitively because motifs can be defined in various different ways and there is no universal solution. Therefore, systematic approaches for motif extraction/definition are necessary.

For this purpose, here we present an improved text mining method named Phylogenetic Tree-Based Motif Finding algorithm (PMF). In this method, relevance between text strings is considered, which are defined by the phylogeny of the strings. By statistically associating the motif counts computed via PMF with disease statuses, efficient classification of disease samples based on (microbiota) sequence texts could be achieved. We have simulated the 16S rRNA datasets of pneumonia and dentes cariosus patients with this pipeline, respectively. Compared to previous results [14], our new pipeline shows better classifications. Additionally, the pipeline is suitable for issues with high-dimensional feature spaces.

2. Data and Methods

2.1. Data and Preprocessing. We acquired 16S rRNA sequencing fasta files of pneumonia patients and dental decay patients from Zhou et al. [2] and Ling et al. [3], respectively. Two to six length k-mer counting results in each meta-genomic sequence were calculated [15]. The k-mer counting results were shown in Files S1 and S2 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/6598307. Each counting and its antisense complementary result were summarized and combined together. The k-mer frequencies were normalized by the reciprocal of length of each sequence as weight and divided by the number of sequences in each fasta file. Identified microbes of 16S rRNAs' sequences from Zhou et al. [2], which were downloaded from NCBI website (ID: GU737566 to GU737625 and HQ914698 to HQ914775) (http://www.ncbi.nlm.nih.gov), were used for constructing the phylogenetic trees. After removing redundant sequences, a total of 90 microbe species were used for further analysis.

The pneumonia samples included 101 patients with hospital-acquired pneumonia (HAP), 43 patients with community-acquired pneumonia (CAP), and 42 normal persons as control. 71 HAP cases, 32 CAP cases, and 30 cases of normal samples were allocated as training data; fitness was calculated using 5-fold proportional cross validation. The other 30 cases of HAP, 13 cases of CAP, and 12 cases of normal were set as the test data, so that classifications were evaluated. For the k-mer counting profiles of 16S rRNAs fasta file collected from dental plaques samples, the training data contained 23 dental decay patients and 20 normal samples and the test data contained 9 dental decay patients and 8 normal samples. For the k-mer counting of 16S rRNAs collected from saliva samples, the training data contained 23 dental decay patient samples and 19 normal samples; and the test data contained 10 dental decay patient samples and 8 normal samples. The partition of the training and test datasets, as well as the cross validation of training data themselves, was adopted from the previous study; hence impartial comparisons (with previous results [14]) could be performed.

Table 1: Alphabet of generalized letters.

Letter	Members	Antisense complementary letter
R	AG	Y
Y	CT	R
W	AT	W
M	AC	K
K	GT	M
S	CG	S
H	ACT	D
B	CGT	V
V	ACG	B
D	CGT	H
N	ACGT	N

2.2. Phylogenetic Tree-Based Motif Finding (PMF) Method. The improved text mining method, Phylogenetic Tree-Based Motif Finding algorithm (PMF), handled counting results of each person's 16S rRNA fasta file. PMF algorithm consisted of three parts: motif finding, motif sorting, and model evaluation. Motif finding was the main part of the algorithm, in which key step was constructing a clustering tree to combine the original strings to a new motif, that is, transforming original letters ("A," "T," "C," and "G") into the generalized letters ("Y," "R," "W," "K," "M," "S," "D," "V," "B," "H," and "N"). The rules of the generalized letters were shown in Table 1.

Minimum distance method was used to cluster the phylogenetic tree. For each pair of sequences with the same length, the phylogenetic distance was calculated by summarizing differences of all sites. For the generalized letters, the differences were calculated using the number of intersections divided by the number of unions. The phylogenetic distance of two motifs was estimated by summarizing differences of both original and generalized sites. If the phylogenetic distance of antisense complementary sequence was smaller than the original sequence, the instance of its antisense complementary sequence was selected. To calculate the complementary generalized letters, each member of the generalized letters was calculated, (i.e., "A" versus "T" and "C" versus "G"),

and results were summarized by rules of Table 1. If there was more than one pair of sequences with the minimum distance, the phylogenetic distances were sorted using Kruskal-Wallis statistics in descending order as follows:

$$\mathrm{KW} = 1 - p^{\mathrm{new}} - \frac{\sum_{i=1}^{n}\left(1 - p_i^{\mathrm{original}}\right)}{n}, \quad (1)$$

where p^{new} was the Kruskal-Wallis test p value of new motif profile, p_i^{original} was the Kruskal-Wallis test p value of ith original sequence covered by the new motif, and n was the number of original sequences or their antisense complementary sequences covered by the new motif. The generalized motif (and its antisense complementary motif) was composed of a group of original sequences; the original sequences with profiling were defined as covered by the new motifs. Profile of the new motif was calculated as follows:

$$\mathrm{profiling\_motif} = \mathrm{profile}\,(:, \mathrm{covered}) * \mathrm{LDA\_weights}, \quad (2)$$

where "profile(:, covered)" were the profiles of original sequences covered by the new motif (i.e., rows were samples and columns were the covered original sequences), and LDA_weights were the linear combination weights calculated by Linear Discriminant Analysis (LDA) [16] method with maximum Fisher's Rayleigh quotient (shown in "Linear Discriminant Analysis" part in the other method). Therefore profiles of the covered sequences were replaced by the profile of new motif.

With the clustering rule, the original sequences could be transformed into the generalized motifs. To suit for high-dimension characteristics of text data, first mth nonredundant pairs of sequences were combined to new motifs, where m = square root (Sqrt in short) of the total number of sequences with the same length. The batch computing method could accelerate motif finding part of PMF and avoid overfitting the training data [17]. Therefore the generalized motifs were found iteratively until all sites changed into "N."

After finding motifs with the same length, different length (e.g., from two to six) motifs needed to be sorted using motif sorting part of PMF by integrating Kruskal-Wallis p value [18] and specificity in descending order. For each motif, the specificity was calculated as follows:

$$\mathrm{specificity} = \frac{\sum_{i=1}^{K}\left(1 - x_i/4\right)}{K}, \quad (3)$$

where K was the length of each sequence and x_i was the number of members of ith site's (generalized) letter. Therefore each motif could be sorted in descending order as follows:

$$1 - p \text{ value} + \text{specificity}. \quad (4)$$

With suitable motifs, original profiles could be merged into profiles of motifs by (2), and covered profiles could be deleted so dimensions reduction would be performed. The more original sequences were replaced by generalized motifs; the linear bias was getting greater, but variance was getting lower, and vice versa. To compromise between the bias and variance criteria, model evaluation part was performed to select necessary motifs. Other than training errors calculated by (5-fold proportional) cross validation, number of dimensions was also considered. Therefore, at most the first p (p = Sqrt$\{N\}$) models with minimum training errors were set as candidate models to be evaluated and combined with dimensions (in descending order). The number of dimensions was considered as the logarithm penalty [17], together with the training errors, so the minimum value model was selected as follows:

$$i \longleftarrow \arg\min \left\{\mathrm{training\_error} + \log(\mathrm{dimension}) * \frac{\log(N * k)}{(N * k)}\right\}, \quad (5)$$

$$N = \min\left\{\mathrm{number\_training\_data}, b\right\}, \quad (6)$$

$$k = \frac{\min(\mathrm{dimension})}{\max(\mathrm{dimension})}, \quad (7)$$

where b was the number of candidate models with lower dimensions than the model minimum training error.

The pipeline of the proposed algorithm is described in detail below and its flowchart was shown in Figure 1.

Step 1 (initialization). Delete features (sequences) with zeros variance profile. Set counters of 2 to 6 length sequences to zero.

Step 2. Enter each loop from Step 3 to Step 5 until the counter reaching $n_k - 1$ for length of sequences is from 2 to 6, respectively (i.e., $n = 2, 3, 4, 5, 6$).

Step 3. Select and sort first $kn/2$ pairs of sequences with minimum phylogenetic distance combined with profile Kruskal-Wallis statistics in descending order as (1).

Step 4 (remove redundancy of the selected pairs). For each current sorted pairs, delete any later selected pairs having intersection with the current one. Finally select $m = \mathrm{Sqrt}(kn)$ pairs at most.

Step 5. Merge the final selected pairs of sequences into the generalized sequences/motifs. Combine profiles of original sequences covered by the generalized motif using Linear Discriminant Analysis (LDA) method with maximum Fisher's Rayleigh quotient value. Counter ← Counter + m.

Step 6. Sort the motifs by the specificity and Kruskal-Wallis p value in descending order using (4).

Step 7. Evaluate models by (5). Treat original profiles by selected suitable motifs using (2).

2.3. Other Methods

2.3.1. Kruskal-Wallis Test. Kruskal-Wallis [18] test is a nonparametric method for testing whether samples originate from the same distribution. The test assumes that all samples from the same group have the same continuous distribution,

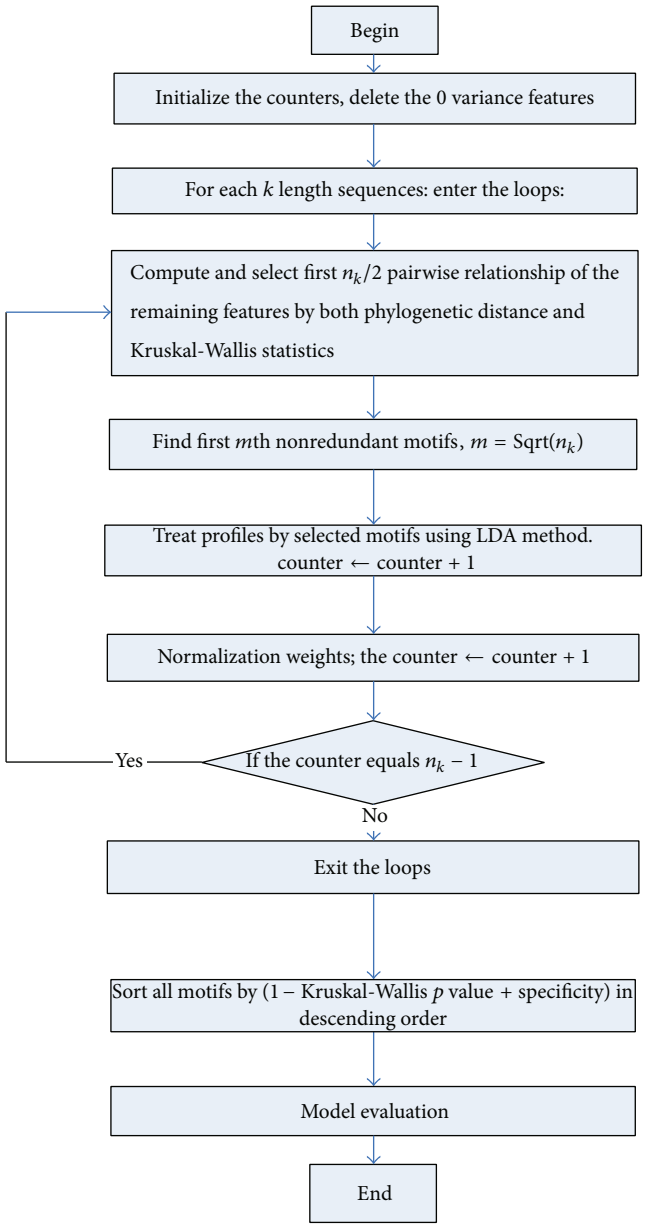


Figure 1: PMF algorithm flowchart.

and they are mutually independent. In this study, Kruskal-Wallis p value was used to rank features.

2.4. Information Gain Method. Information Gain [19] measures the classification ability of each feature with respect to the relevance with the output class, which is defined as Information Gain = $H(S) - H(S \mid x)$:

$$H(S) = -\sum_{s \in S} p(s) \log_2 (p(s)),$$
$$H(S \mid x) = -\sum_{x \in X} p(x) \sum_{s \in S} p(s \mid x) \log_2 (p(s \mid x)), \quad (8)$$

where S and x are features. When measuring the mutual relation between the extracted features and the class, Information Gain is also known as mutual information. k-mer counting values were discretized using two thresholds' mean ± std. If more than one sequence was with the same Information Gain value, they were sorted by Kruskal-Wallis p value.

2.5. Chi-Square Statistic. This method uses the Chi-square statistic to discretize numeric attributes and achieves feature selection via discretization [20]. The Chi-square value is defined as

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$
$$E_{ij} = \frac{M_i B_j}{N}, \quad (9)$$

where c is the number of intervals, k is the number of classes, A_{ij} is the number of samples in the ith interval and the jth class, M_i is the number of samples in the ith interval, B_j is the number of samples in the jth class, and N is the total number of samples. k-mer counting values were discretized using two thresholds' mean ± std. If more than one sequence was with the same Chi-square statistic, they were sorted by Kruskal-Wallis p value.

2.6. Linear Discriminant Analysis. Linear Discriminant Analysis (LDA) is a typical variable transformation method to reduce dimensions [16]. The key step of LDA is to maximize the Rayleigh quotient:

$$J(W) = \frac{\alpha^T S_B \alpha}{\alpha^T S_W \alpha}, \quad (10)$$

where the "between-class scatter matrix" is defined as

$$S_B = \sum_k (p_k - 1) \sum \frac{(m_k - m)(m_k - m)'}{(K - 1)} \quad (11)$$

and the "within-class scatter matrix" is defined as

$$S_W = \sum_k^K \frac{(y - m_k)(y - m_k)'}{(N - K)}. \quad (12)$$

K is the number of classes, p_k is the number of the samples within the kth class, m_k is the mean value of the sample within the kth class, and m is the mean value of all the samples.

Traditional LDA requires the total scatter matrix to be nonsingular. To deal with the singularity problems, classical LDA method was modified in a way that a unit diagonal matrix with small weights was added to the within-class scatter matrix, if the scatter matrix is singular [14].

3. Results and Discussion

We first performed the PMF algorithm on pneumonia samples. We considered both 2-class problem (pneumonia: CAP + HAP, versus normal) and 3-class problem (HAP, CAP, versus normal). Conventionally, due to the data imbalance,

Table 2: Classification results of pneumonia data in 3-class problem.

Method	Error rate		Dimension	Feature
	On training data	On test data		
SVM/FMS	0.1895	0.2637	29	Microbes
SVM/PMF	0.062	0.0756	411	Sequences
SVM/Kruskal-Wallis	0.1187	0.5273	272	Sequences
SVM/Information Gain	0.143	0.2124	12	Sequences
SVM/Chi-square statistic	0.1743	0.5909	1280	Sequences
SVM	0.2187	0.3812	4390	Sequences
NNA/FMS	0.2013	0.3406	112	Microbes
NNA/PMF	0.2152	0.2081	786	Sequences
NNA/Kruskal-Wallis	0.2718	0.3363	85	Sequences
NNA/Information Gain	0.2354	0.4141	39	Sequences
NNA/Chi-square statistic	0.2649	0.3107	69	Sequences
NNA	0.442	0.6162	4390	Sequences

the accuracy for each class was used to measure the classification, which was equivalent to combining the specificity and sensitivity in general classifications. Two widely applied methods, nearest neighbor algorithm (NNA) and support vector machine (SVM), were used to select the optimal classifier set of motifs extracted by PMF for pneumonia samples. Since SVM mainly suits pairwise classifications, normal samples must be discriminated against the pneumonia samples (CAP and HAP) before CAP and HAP were classified in a 3-class problem. To evaluate the performance of our (k-mer) motif-oriented method, we compared our results with those of previous methods, including the Feature Merging and Selection algorithm (FMS) based on sequence-microbe associations [14], as well as other k-mer counting feature selection algorithms, for example, the Information Gain method [19], Chi-square statistic method [20], and primitive Kruskal-Wallis statistic method [18].

Figure 2 showed the learning curves for the training data; combined with logarithm penalty evaluation, the best evaluated models were selected with 1321 and 1369 runs of PMF for the 3-class problem, with SVM and NNA classifiers, respectively. Optimal models for the 2-class problem were selected with 1218 and 1136 runs of PMF (with SVM and NNA). As shown in Tables 2 and 3, our method had the lowest mean error in both 3-class and 2-class problem (with either SVM or NNA combined), compared with the previous methods mentioned earlier.

In statistics, a receiver operating characteristic (ROC) curve is the summary of both sensitivity and specificity for various thresholds. ROC was constructed for each subset of features (Figure 3). As shown, the optimal features that are selected under the combined criteria of cross validation and model evaluation possessed high specificity (~80%) with high sensitivity (~70%) for the 3-class problem, indicating the ability of our method. Moreover, even higher specificity and sensitivity were obtained (>0.95) for the 2-class problem. Noteworthy, PMF combined with SVM performed better in the classification; therefore the results derived by PMF with SVM for the k-mer counting profiles of pneumonia samples were used for further analysis.

Heat map is a frequently used matrix of pairwise sample correlations indicating anticorrelation or correlation using a color scale, that is, green to red. Figure 4(a) showed that the original data profile was almost invisible for patterns or sample classifications, after being analyzed by our method, since the original feature space had been reduced to a much smaller space spanned by a few features (with the most important variances retained). Therefore as shown in Figures 4(b) and 4(c), the heat maps of the samples were much clearer with high resolutions for classifications.

Profiles with reduced dimensions obtained by PMF were sorted according to the Kruskal-Wallis p values. The top 5 motifs with p value < 0.05 in 3-class problem were "KCTCWT," "TTCGHT," "CGATCS," "TCWCTA," and "TTWCGC". Sequences (including antisense complementary sequences) covered by the first motif "KCTCWT" (p value = 0.0126) were matched to the microbe taxonomic results of Zhou et al. [2]. 6 out of 19 matched microbes were among the top 20 genera suspiciously contributing to pneumonia [2] (Table S1). 10 out of the remaining 13 microbes were also related to pneumonia [21–30] (Table S1). The 2 motifs with p value < 0.05 in 2-class problem were "WTCGTC" and "ATCWCT". Sequences covered by first one "WTCGTC" (p value = 0.0288) were matched to the published taxonomic data [2]. Five out of 6 matched microbes were related to pneumonia [2, 29, 30] (Table S2). Furthermore, by pinpointing the 25 (e.g., 19 + 6) matched microbes in a phylogenetic tree constructed from the published taxonomic data (using MEGA6 software [31] with minimum distance method), we observed that distribution of the microbes was dispersed, indicating the diverse functions performed by microbiota for human (Figure 5).

Our method was also tested on k-mer counting profiles from dental decay sample. These samples were collected from saliva and dental plaques separately. Combined with logarithm penalty evaluation, the best evaluated models were

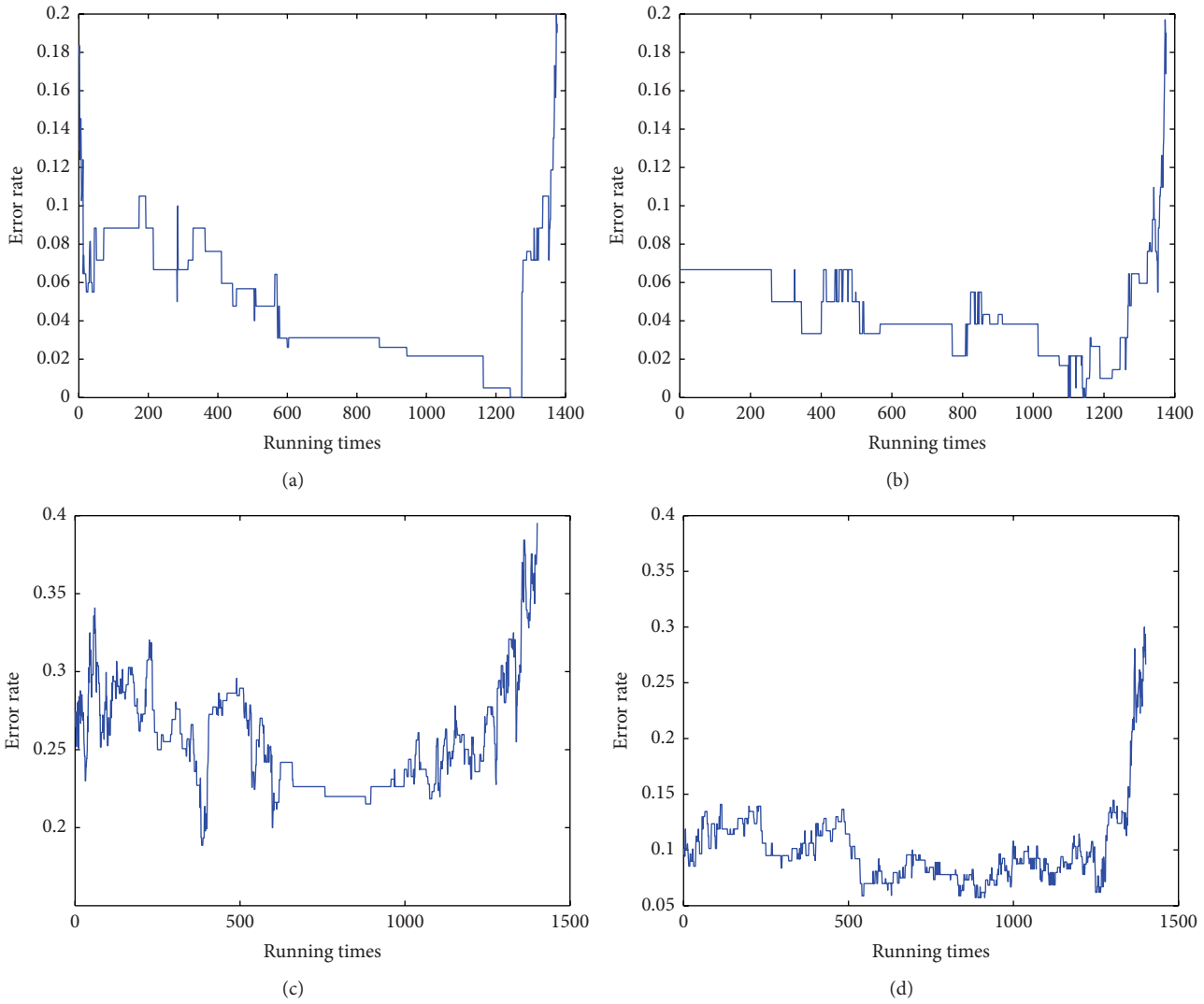
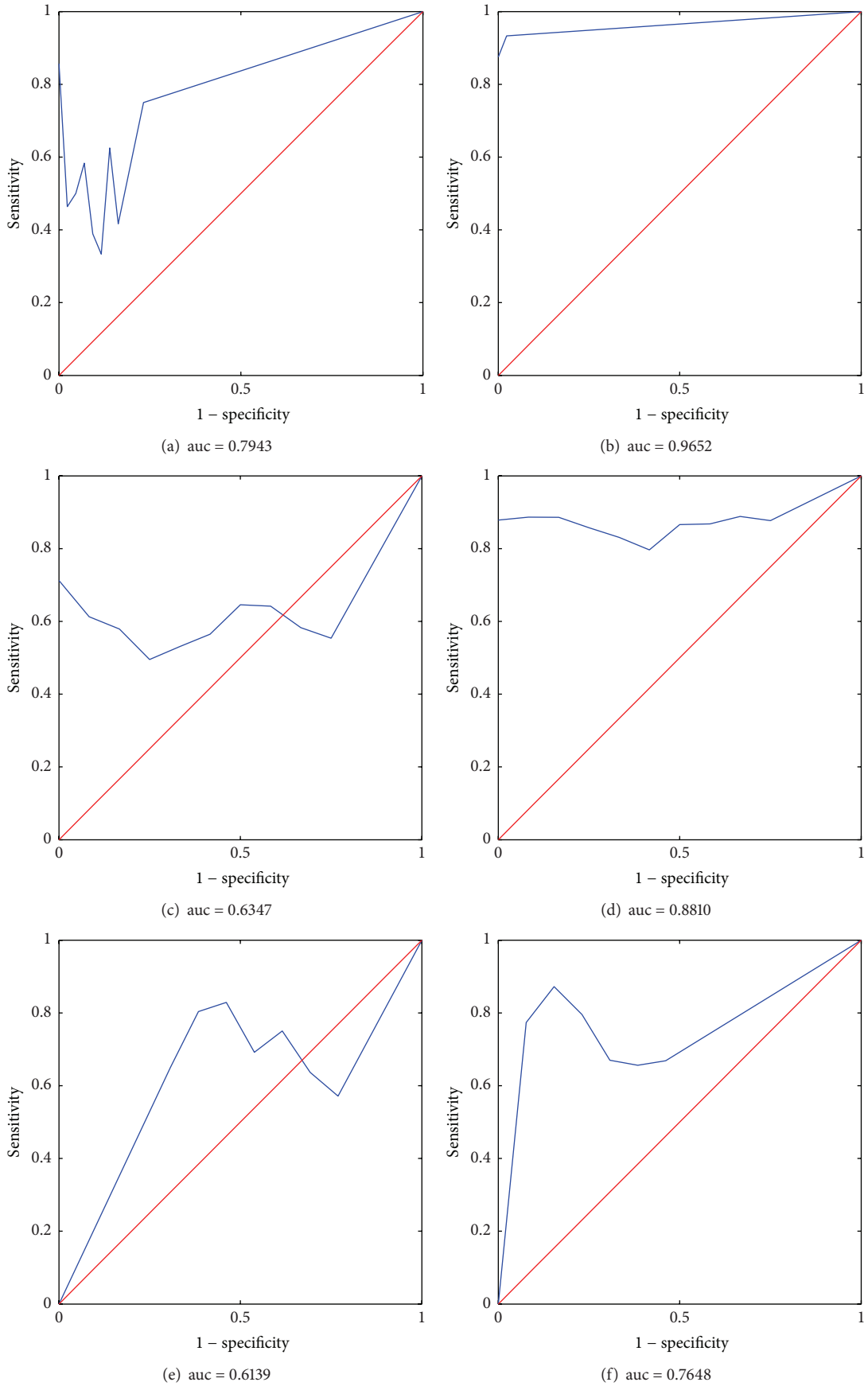


Figure 2: Learning curves of PMF algorithm for 3-class problem with NNA (a), for 3-class problem with SVM (b), for 2-class problem with NNA (c), and for 2-class problem with SVM (d).

Table 3: Classification results of pneumonia data in 2-class problem.

Method	Error rate		Dimension	Feature
	On training data	On test data		
svm/FMS	0.0922	0.1279	42	Microbes
svm/PMF	0	0	551	Sequences
svm/Kruskal-Wallis	0.01	0	28	Sequences
svm/Information Gain	0	0.0116	26	Sequences
svm/Chi-square statistic	0.01	0.0417	127	Sequences
svm	0.0667	0.0116	4390	Sequences
NNA/FMS	0.1279	0.2393	20	Microbes
NNA/PMF	0	0.0833	361	Sequences
NNA/Kruskal-Wallis	0.0167	0.125	13	Sequences
NNA/Information Gain	0.0214	0.125	12	Sequences
NNA/Chi-square statistic	0.0381	0.125	26	Sequences
NNA	0.2667	0.5	4390	Sequences



(a) auc = 0.7943

(b) auc = 0.9652

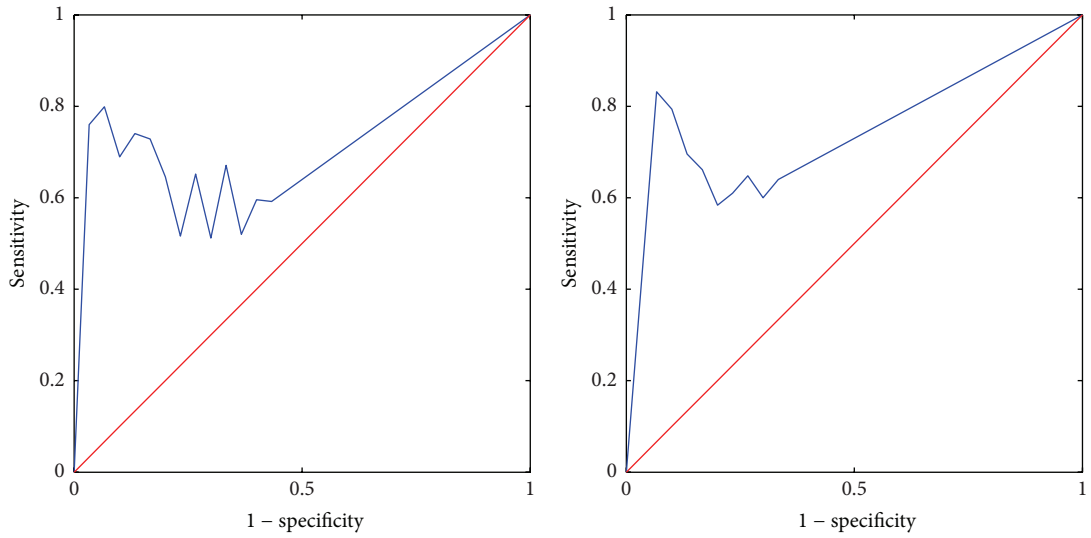
(c) auc = 0.6347

(d) auc = 0.8810

(e) auc = 0.6139

(f) auc = 0.7648

Figure 3: Continued.



(g) auc = 0.7219 (h) auc = 0.7520

FIGURE 3: ROC curves of motif finding step of PMF algorithm for pneumonia samples (CAP + HAP) in 2-class problem (a, b), normal samples (c, d), CAP samples in 3-class problem (e, f), and HAP samples in 3-class problem (g, h), with NNA or SVM.

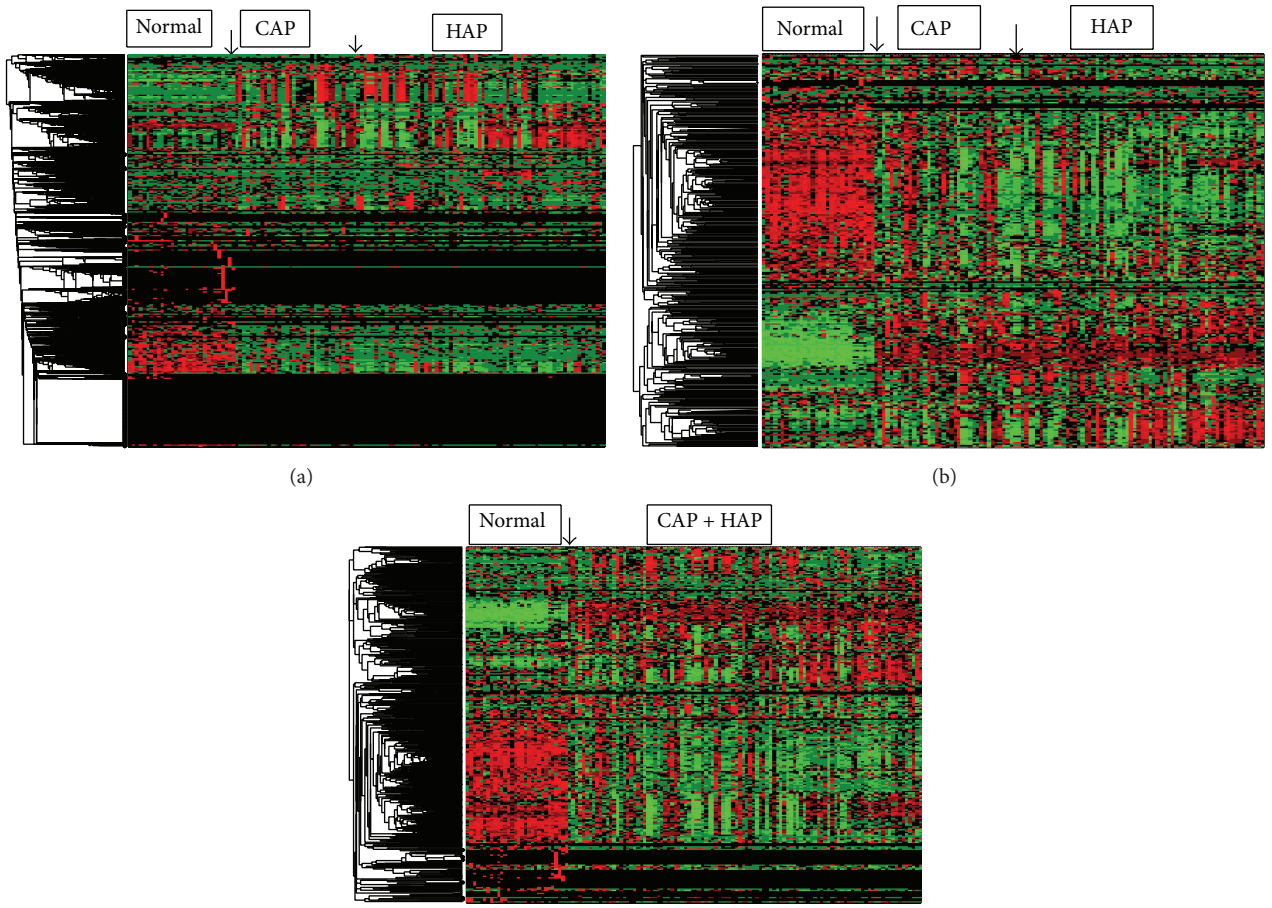


FIGURE 4: Heat map of k-mer counting profiles of original pneumonia data for 3-class problem (a), data after treating by PMF for 3-class problem, (b) and data after treating by PMF for 2-class problem (c). Rows are retained motifs and columns are disease classes. From left to right are 30 normal, 32 CAP, and 71 HAP samples for 3-class problem and 30 normal 103 pneumonia samples for 2-class problem.

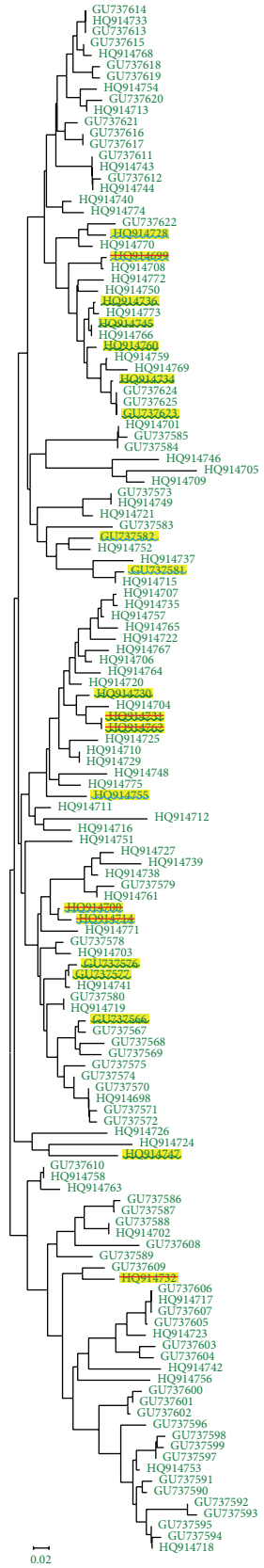


Figure 5: Phylogenetic relationship of identified microbiota signatures. Identified microbes matched by significant motifs are highlighted with underline ("KCTCWT") or strikethrough ("WTCGTC").

selected with 2330 and 2485 runs of PMF for dental plaques (with SVM) and saliva samples (with NNA), respectively (Figure S1). The results showed that our method could also select suitable classifiers and perform better on the test data than the previous and other methods (Tables S3 and S4).

4. Conclusions

In this paper, we presented the PMF method to analyze the align-free k-mer counting profiles of 16S rRNA microbial data. The improved pipeline systematically analyzed relevance between each pair of sequences using minimum distance phylogenetic trees. Moreover, PMF also considered relationships between k-mer counting profiles and the disease status. In addition, by combining original profiles using the LDA method, PMF learned profiles of text strings suitable for disease classification. Batching method also accelerated PMF and avoided overfitting of training data. As a result, via combing characteristics of sequences and classification statistics of text profiles, PMF selected suitable motifs to evaluate metagenome characteristics of human microbiota disease.

In conclusion, we developed an improved motif-based text mining algorithm, and the new pipeline was verified by both pneumonia and dentes cariosus samples. As the classification results have shown, it was demonstrated that PMF was an effective approach for finding informative motifs from training data, and it was validated well compared with the previous study and other widely used methods. PMF performed well and it could extend evolutionary/phylogenetic approaches to perform supervised learning on microbiota text data to discriminate disease/pathology status.

Abbreviations

PMF: Phylogenetic Tree-Based Motif Finding algorithm
FMS: Feature Merging and Selection algorithm
LDA: Linear Discriminant Analysis
HAP: Hospital-acquired pneumonia
CAP: Community-acquired pneumonia
NNA: Nearest neighbor algorithm
SVM: Support vector machine.

Disclosure

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

Conflict of Interests

The authors declared that they have no competing interests.

Authors' Contribution

Yin Wang performed algorithm design and wrote the paper. Yuhua Zhou and Zongxin Ling collected the data. Lei Liu, Lu Xie, and Xiaokui Guo designed and sponsored the study. Rudong Li contributed and edited the paper. All authors read and approved the paper. Yin Wang and Rudong Li equally contributed to this paper.

Acknowledgments

This work was supported by the National High Technology Research and Development Program (863 Program) (2012AA02A602, 2015AA020104), National Science and Technology Major Project (2012ZX09303013-015), and special funds for scientific research in the health industry (201302010).

References

[1] H. Sokol, B. Pigneur, L. Watterlot et al., "*Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients," *Proceedings of the National Academy of Sciences of the United States*, vol. 105, no. 43, pp. 16731–16736, 2008.

[2] Y. Zhou, P. Lin, Q. Li et al., "Analysis of the microbiota of sputum samples from patients with lower respiratory tract infections," *Acta Biochimica et Biophysica Sinica*, vol. 42, no. 10, pp. 754–761, 2010.

[3] Z. Ling, J. Kong, P. Jia et al., "Analysis of oral microbiota in children with dental caries by PCR-DGGE and barcoded pyrosequencing," *Microbial Ecology*, vol. 60, no. 3, pp. 677–690, 2010.

[4] Z. Gao, G. I. Perez-Perez, Y. Chen, and M. J. Blaser, "Quantitation of major human cutaneous bacterial and fungal populations," *Journal of Clinical Microbiology*, vol. 48, no. 10, pp. 3575–3581, 2010.

[5] E. M. Bik, P. B. Eckburg, S. R. Gill et al., "Molecular analysis of the bacterial microbiota in the human stomach," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 3, pp. 732–737, 2006.

[6] D. Knights, E. K. Costello, and R. Knight, "Supervised classification of human microbiota," *FEMS Microbiology Reviews*, vol. 35, no. 2, pp. 343–359, 2011.

[7] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Applied and Environmental Microbiology*, vol. 73, no. 16, pp. 5261–5267, 2007.

[8] J. Handelsman, "Metagenomics: application of genomics to uncultured microorganisms," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 4, pp. 669–685, 2004.

[9] C. Simon and R. Daniel, "Metagenomic analyses: past and future trends," *Applied and Environmental Microbiology*, vol. 77, no. 4, pp. 1153–1161, 2011.

[10] S. Karlin, J. Mrázek, and A. M. Campbell, "Compositional biases of bacterial genomes and evolutionary implications," *Journal of Bacteriology*, vol. 179, no. 12, pp. 3899–3913, 1997.

[11] T. Thomas, J. Gilbert, and F. Meyer, "Metagenomics—a guide from sampling to data analysis," *Microbial Informatics and Experimentation*, vol. 2, no. 1, article 3, 2012.

[12] T.-H. Lin, R. F. Murphy, and Z. Bar-Joseph, "Discriminative motif finding for predicting protein subcellular localization," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 2, pp. 441–451, 2011.

[13] J. Korhonen, P. Martinmäki, C. Pizzi, P. Rastas, and E. Ukkonen, "MOODS: fast search for position weight matrix matches in

DNA sequences," *Bioinformatics*, vol. 25, no. 23, pp. 3181–3182, 2009.

[14] Y. Wang, Y. Zhou, Y. Li et al., "An improved dimensionality reduction method for meta-transcriptome indexing based diseases classification," *BMC Systems Biology*, vol. 6, supplement 3, article S12, 2012.

[15] F. Zhou, V. Olman, and Y. Xu, "Barcodes for genomes and applications," *BMC Bioinformatics*, vol. 9, article 546, 2008.

[16] X.-Y. Jing, D. Zhang, and Y.-Y. Tang, "An improved LDA approach," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 5, pp. 1942–1951, 2004.

[17] A. E. Isabelle Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[18] L. J. Wei, "Asymptotic conservativeness and efficiency of kruskal-wallis test for K dependent samples," *Journal of the American Statistical Association*, vol. 76, no. 376, pp. 1006–1009, 1981.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

[20] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," in *Proceedings of the 7th International Conference on Tools with Artificial Intelligence*, pp. 388–391, Herndon, Va, USA, November 1995.

[21] N. Jafari, R. Behroozi, D. Farajzadeh, M. Farsi, and K. Akbari-Noghabi, "Antibacterial activity of *Pseudonocardia* sp. JB05, a rare salty soil actinomycete against *Staphylococcus aureus*," *BioMed Research International*, vol. 2014, Article ID 182945, 7 pages, 2014.

[22] H. Lu, G. Qian, Z. Ren et al., "Alterations of *Bacteroides* sp., *Neisseria* sp., *Actinomyces* sp., and *Streptococcus* sp. populations in the oropharyngeal microbiome are associated with liver cirrhosis and pneumonia," *BMC Infectious Diseases*, vol. 15, no. 1, article 239, 2015.

[23] P. Li, C. Yang, J. Xie et al., "*Acinetobacter calcoaceticus* from a fatal case of pneumonia harboring blaNDM-1 on a widely distributed plasmid," *BMC Infectious Diseases*, vol. 15, article 131, 2015.

[24] A. Hasegawa, T. Sato, Y. Hoshikawa et al., "Detection and identification of oral anaerobes in intraoperative bronchial fluids of patients with pulmonary carcinoma," *Microbiology and Immunology*, vol. 58, no. 7, pp. 375–381, 2014.

[25] M. M. Pettigrew, A. S. Laufer, J. F. Gent, Y. Kong, K. P. Fennie, and J. P. Metlay, "Upper respiratory tract microbial communities, acute otitis media pathogens, and antibiotic use in healthy and sick children," *Applied and Environmental Microbiology*, vol. 78, no. 17, pp. 6262–6270, 2012.

[26] J. P. Balikian, P. G. Herman, and J. J. Godleski, "Serratia pneumonia," *Radiology*, vol. 137, no. 2, pp. 309–311, 1980.

[27] L. Yu, A. H. Gunasekera, J. Mack et al., "Solution structure and function of a conserved protein SP14.3 Encoded by an essential *Streptococcus pneumoniae* gene," *Journal of Molecular Biology*, vol. 311, no. 3, pp. 593–604, 2001.

[28] P. Panagou, L. Papandreou, and D. Bouros, "Severe anaerobic necrotizing pneumonia complicated by pyopneumothorax and anaerobic monoarthritis due to *Peptostreptococcus magnus*," *Respiration*, vol. 58, no. 3-4, pp. 223–225, 1991.

[29] C. M. Chao, C. C. Lai, H. Y. Tsai et al., "Pneumonia caused by *Aeromonas* species in Taiwan, 2004–2011," *European Journal of Clinical Microbiology and Infectious Diseases*, vol. 32, no. 8, pp. 1069–1075, 2013.

[30] M. Koide, F. Higa, M. Tateyama, H. L. Cash, A. Hokama, and J. Fujita, "Role of *Brevundimonas vesicularis* in supporting the growth of *Legionella* in nutrient-poor environments," *New Microbiologica*, vol. 37, no. 1, pp. 33–39, 2014.

[31] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "MEGA6: molecular evolutionary genetics analysis version 6.0," *Molecular Biology and Evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.