Naïve Bayesian Classifiers with Multinomial Models for rRNA Taxonomic Assignment

Kuan-Liang Liu and Tzu-Tsung Wong

Abstract—The introduction of next-generation sequencing in ecological studies has created a major revolution in microbial and fungal ecology. Direct sequencing of hypervariable regions from ribosomal RNA genes can provide rapid and inexpensive analysis for ecological communities. To get deep understanding from these rRNA fragments, the Ribosomal Database Project developed the "RDP Classifier" utilizing 8-mer nucleotide frequencies with Bayesian theorem to obtain taxonomy affiliation. The classifier is computationally efficient and works well with massive short sequences. However, the binary model employed in the RDP classifier does not consider the repetitive 8-mers in each reference sequence. Previous studies have pointed out that multinomial model usually results a better performance than binary model. In this study, we present the naïve Bayesian classifiers with multinomial models that take repetitive 8-mers into account for classifying microbial 16S and fungal 28S rRNA sequences. The results obtained from the multinomial approach were compared with those obtained from the binomial RDP classifier by 250-bp, 400-bp, 800-bp, and full-length reads to demonstrate that the multinomial approach can generally achieve a higher predictive accuracy in most hypervariable regions.

Index Terms—Naïve Bayesian classifier, taxonomy assignment, rRNA

1 Introduction

PHYLOGENETIC analyses of ribosomal RNA (rRNA) gene have been extensively carried out to identify bacterial species and perform taxonomic studies [10], [24]. The power of rRNA for phylogenetic analysis can be attributed to several factors, including its presence in all cellular organisms, its favorable patterns of sequence conservation that enable study of evolutionary events, and the ease with which this gene can be cloned and sequenced from new organisms [18]. The ability to PCR-amplify (hundreds of) thousands of rRNA genes in a single reaction has revolutionized our understanding of microbial diversity in nature and has allowed the effective profiling of complex communities, including various environmental and human microbiomes [2], [6], [8], [18], [20], [21].

The 16S rRNA genes contain nine hypervariable regions that have been exploited for the identification and phylogenetic analysis in prokaryotic studies [22]. The combination of highly variable and conserved regions facilitates primer design and provides bar-coding signatures that are useful for the taxonomic placement of individual species as well as groups of related species. Similarly, D1 and D2 variable regions from fungal large-subunit rRNA (LSU) gene have also been commonly used to study fungal diversity [11], [16]. Recently, advances in high-throughput sequencing techniques, such as barcoded pyrosequencing and Illumina sequencing, have substantially increased the depth to which 16S and 28S rRNA genes can be surveyed in complex environmental samples [1], [3], [7], [9], [12]. However, these next-generation sequencing (NGS) technologies typically generate short sequence reads, which raise challenges in rapid and reliable taxonomic assignment. Consequently, effective and efficient classification tools should be developed such that the technologies about massive data set generation and accurate annotation can be useful for accomplishing ecological studies.

The Ribosomal Database Project [4] used extensive databases of sequenced 16S rRNA genes to develop a very robust RDP classifier for 16S rRNA sequences [23]. This classifier is alignment-free, very rapid, and yields accurate results. It has been utilized in various projects to examine the diversity of microbial and fungal community diversities [3], [7], [9], [26]. Recently, this classifier has been applied to multiple taxonomic schemes that are useful for addressing current incongruencies in taxonomic nomenclature among fungal curators [13].

The RDP classifier extracts 8-mer nucleotides from query sequence read by the sliding window method and includes every one of them in calculating classification probabilities. However, in the training phase, the probability of occurrence of each 8-mer nucleotide given a class value is estimated by the binary occurrence of 8-mer nucleotides from the training data. Rosen et al. [19] proposed an NBC classifier without smoothing for classifying whole-genome shotgun metagenomic reads. This classifier uses the same method as the RDP method for extracting features from the testing sequence read. However, during the training phase, NBC classifier uses a multinomial model and counts the frequencies of 8-mer nucleotides to estimate occurrence probabilities. The NBC classifier has been used accurately to identify metagenomic reads, but no previous studies has applied it for the taxonomic assignment of rRNA gene sequences.

This paper aims to improve the performance of naive Bayesian classifiers by taking into account every occurrence of a k-mer nucleotide in the training phase. It presents an extensive empirical comparison between binomial and multinomial models on 16S and 28S gene sequence data sets. A sliding window for 250-bp, 400-bp, and 800-bp short read fragments extraction is also applied to investigate the performance of both models.

The remainder of this paper is organized as follows: Section 2 introduces the 16S and 28S rRNA sequence data sets and presents both the binomial and the multinomial models of the naive Bayesian classifier. The use of a flattening constant to improve classification accuracy is also addressed. Section 3 presents experimental results obtained using the proposed multinomial approach and the RDP classifier. These include comparisons of full-length and short read with various flattening constants and they demonstrate that the multinomial method generally performs better. The final section will discuss the contribution of this work and make suggestions regarding directions for further investigation.

2 Materials and Methods

This section first describes the 16S and 28S sequence data sets for experimental demonstration and introduces the probabilistic framework of the naïve Bayesian classifier. Then, the binomial and the multinomial models for calculating classification probabilities are presented for gene sequence data. The use of flattening constants is a general approach to ensure that every probability estimate for the naïve Bayesian classifier is positive. The purpose of using flattening constants for classifying gene sequence data will also be discussed.

2.1 rRNA Training Set Preparation

A sequence is called a singleton if its class value appears only once in the sequence set. Since classifiers will be evaluated by leave-one-out cross validation (LOOCV) in this study, all singletons were excluded from the original data sets. A total of 7,045 16S rRNA sequences with 1,399 genera were downloaded from the RDP database. Associated taxonomic assignment information was derived from Bergey's *Taxonomic Outline of the Prokaryotes* (release

• *The authors are with the Institute of Information Management, National Cheng Kung University, 1 Ta-Sheuh Road, Tainan City 701, Taiwan.*

Manuscript received 19 Dec. 2012; revised 28 June 2013; accepted 5 Sept. 2013; published online 18 Sept. 2013.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2012-12-0327. Digital Object Identifier no. 10.1109/TCBB.2013.114.

TABLE 1
The Numbers of 16S and 28S rRNA Taxa and Singletons at Different Ranks

	16S SSU rRNA (7045/ 6386)		Fungal 28S LSUrRNA (8506/ 7730)	
	# of taxa	# of singleton	# of taxa	# of singleton
Genera	1399	659	1702	776
Families	224	49	412	108
Orders	98	16	124	16
Classes	44	1	40	8
Phyla	40	5	9	1
Domains	2	0	2	0

5.0 [2004]). According to the class values in the genus rank, this 16S rRNA sequence data set has 6,386 nonsingleton sequences with lengths from 204 to 3,035 bases. Each sequence was labeled with a set of taxa from domain to genus, as defined by Bergey's manual, and the numbers of sequences at each taxa rank were listed in Table 1. These 16S rRNA sequences were then aligned with the *Escherichia coli* reference sequence J01695 using the RDP 10 alignment [5] to generate the master alignment for extracting sequence fragment. A total of 8,506 curated fungal 28S rRNA sequences, spanning the first 1,400-bp of the LSU gene, were provided by Liu et al. [13]. The fungal training data set comprised 1,702 validated fungal genera, spanning 40 classes and 9 phyla, as presented in Table 1. The rest of the 7,730 LSU rRNA nonsingletons were aligned against the reference sequence *Saccharomyces cerevisiae* LSU rRNA (RDN25-1) using ARB Silva [14] to generate the master alignment for sequence fragment extraction.

2.2 Probabilistic Framework of Naïve Bayesian Classifiers for Taxonomic Assignment

The rRNA taxonomic assignment can be viewed as a general sequence read classification problem: sequence reads with known class values were utilized to determine the class value of a new query sequence read, r. The Bayesian classifier assigns the class value C_i that maximizes the posterior probability $P(C_i \mid r)$ to be the predicted class value of the sequence read, r. By Bayes' theorem,

$$P(C_i \mid r) = P(r \mid C_i)P(C_i)/P(r). \tag{1}$$

Since *P(r)* is independent of the class values, it can be ignored in calculating the classification probability. The value of $P(C_i)$ depends on the origin of the sample and is often unknown, so all class values can be reasonably assumed to be equally probable. When $P(r)$ and $P(C_i)$ in (1) are neglected, the taxonomic assignment for sequence read r is

$$\hat{C} = arg\max_i P(C_i \mid r) = arg\max_i P(r \mid C_i). \tag{2}$$

Let a sequence read with n nucleotides $r = (s_1, s_2, \ldots, s_n)$ be represented as a vector of $m = n - k + 1$ overlapping k-mer words $r = (w_1^{(k)}, w_2^{(k)}, \ldots, w_m^{(k)})$, where $w_j^{(k)} = (s_j, s_{j+1}, \ldots, s_{j+k-1})$ for $j = 1, 2, \ldots, m$. The naïve Bayesian classifier assumes that the occurring probabilities of all k-mer words are independent for any given class value. Based on this conditional independence assumption, the probability $P(r \mid C_i)$ can be rewritten as

$$P(r \mid C_i) = \prod_{j=1}^{m} P\big(w_j^{(k)} \mid C_i\big). \tag{3}$$

Since every occurrence of $w_j^{(k)}$ is taken into account in calculating $P(r \mid C_i)$, the model that is used in the testing phase is multinomial. However, both binomial and multinomial models can be adopted to estimate $P(w_j^{(k)} \mid C_i)$ in (3) from the training data, as introduced in the following two sections.

2.3 Binomial Model

Let the number of sequences with class value C_i in the training data be U, and let $u(w_j^{(k)})$ be the number of U sequences that contain the k-mer word $w_j^{(k)}$. The binomial model estimates the probability of $w_j^{(k)}$ given C_i as

$$P\big(w_j^{(k)} \mid C_i\big) = u\big(w_j^{(k)}\big)/U. \tag{4}$$

In this model, every training sequence is distinguished by whether it contains k-mer word $w_j^{(k)}$ or not. When the training data do not include any sequence with class value C_i and $w_j^{(k)}$, the estimate for $P(w_j^{(k)}|C_i)$ will be zero. To solve this problem, probability $P(w_j^{(k)}|C_i)$ is estimated as

$$P\big(w_j^{(k)} \mid C_i\big) = \big[u\big(w_j^{(k)}\big) + Q_j\big]/(U+1), \tag{5}$$

where Q_j is a word-specific constant which ensures that $P(w_j^{(k)} \mid C_i)$ will be positive. Accordingly, the RDP classifier estimates $P(w_j^{(k)}|C_i)$ and uses the Jeffreys-Perks law of succession to determine the word-specific constant as

$$Q_j = \big[v\big(w_j^{(k)}\big) + 0.5\big]/(V+1), \tag{6}$$

where V denotes the total number of training sequences and $v(w_j^{(k)})$ is the number of training sequences that contain $w_j^{(k)}$. Since $Q_j = 1/2$ when both V and $v(w_j^{(k)})$ are zero, the binomial model is again used to calculate the word-specific constant.

2.4 Multinomial Model

In contrast to the binomial model, the multinomial model considers frequencies of k-mer words to estimate $P(w_j^{(k)}|C_i)$ in the training phase as

$$P\big(w_j^{(k)} \mid C_i\big) = f\big(w_j^{(k)} \mid C_i\big)/a(C_i), \tag{7}$$

TABLE 2
The Classification Accuracies of Various Models for Full-Length rRNA Sequences in Rank Genus

	MM-1	MM-0.1	MM-0.01	MM-0.001	MM-0.0001	MM-0.00001	MM-0.000001	BM
16S rRNA	66.05%	88.18%	95.30%	95.51%	**95.52%**	95.24%	95.00%	94.86%
28S rRNA	46.58%	70.15%	80.65%	81.80%	82.03%	**82.12%**	82.00%	78.65%

where $f(w_j^{(k)} \mid C_i)$ is the frequency of $w_j^{(k)}$ in the training sequences with class value C_i, and $a(C_i)$ is the total number of k-mer words in the training sequences. Sometimes, no training sequences will include a k-mer word $w_j^{(k)}$ that is extracted from a testing read, so $f(w_j^{(k)} \mid C_i) = 0$. In this case, the result of classifying the testing read is likely to be distorted.

Since every nucleotide in a k-mer word must be a, t, c, or g, the number of possible k-mer words is 4^k. When data are not observed, the probabilities of occurrence of the 4^k k-mer words should be equally likely, and this is called noninformative assumption. To ensure that every probability estimate that is calculated in the multinomial model is positive, the m-estimate approach proposed by Mitchell [15] will be used to calculate $P(w_j^{(k)} \mid C_i)$ as

$$P(w_j^{(k)} \mid C_i) = \left[f(w_j^{(k)} \mid C_i) + \alpha^c\right]/(\mathrm{a}(C_i) + \alpha^c \times 4^k), \qquad (8)$$

where α^c is a flattening constant that is chosen to represent the confidence level about the noninformative assumption. A larger α^c indicates that all k-mer words are more likely to have the same probability of occurrence. According to the prior concept proposed by Wong [25] for the naïve Bayesian classifier, the confidence level of this noninformative prior equals $\alpha^c \times 4^k$. In the experiment performed in this study, the value of α^c will be set to 1, 0.1, 0.01, 0.001, 0.0001, 0.00001, or 0.000001 for the multinomial model.

3 Experimental Results

This section provides empirical evidence that the naïve Bayesian classifier generally performs better when the multinomial model is used in the training phase to estimate the occurrence probabilities of k-mer words. The comparison will be made not only on full read length, but also for short read fragments with length 250-bp, 400-bp, and 800-bp. As indicated by Wang et al. [23], the naïve Bayesian classifier will have a better performance when the value of k is either eight or nine. For computational efficiency, the value of k is set to 8. Both models were evaluated using the leave-one-out cross validation method. Every sequence in a set is in turn used for testing and the testing result was induced from the other sequences in the set by a classifier. The proportion of the sequences with correct prediction will be the accuracy of the classifier. For simplicity, the binomial model and the multinomial model with flattening constant α^c are denoted by BM and MM-α^c, respectively. Each classifier was written in C++ code and perl script was used to implement the LOOCV process. All source code is available online at https://sourceforge.net/p/gdrrnaclassifie/wiki/Home/.

3.1 Full-Length Comparison

Table 2 presents the LOOCV results concerning various models for full-length sequences in rank genus, where a bold value in a row indicates the most accurate result for the gene sequence set in that row. Table 2 reveals that the MM-0.0001 model has the highest accuracy on 16S rRNA gene sequence set and the MM-0.00001 has the highest accuracy on 28S rRNA gene sequence set. When the flattening constant is less than or equal to 0.01, the resulting accuracy is stable. As noted in Section 2.4, the confidence level in a noninformative prior is $\alpha^c \times 4^k$. Accordingly, the confidence levels for MM-1, MM-0.1, MM-0.01, MM-0.001, MM-0.0001, MM-0.00001, and MM-0.000001 are 65,536, 6,553.6, 655.36, 65.536, 6.5536, 0.65536, and 0.065536, respectively, when $k = 8$. The confidence levels for both MM-1 and MM-0.1 models are too high such that the noninformative assumption plays an important role in classifying a sequence. This is inappropriate because class prediction should be primarily determined by training data. Since the accuracies of the MM-1 and the MM-0.1 models are greatly lower than the others, they were neglected in the short read comparison.

Figs. 1 and 2 display the comparison of MM and BM models by each taxon on the full-length 16S rRNA and the 28S fungal LSU

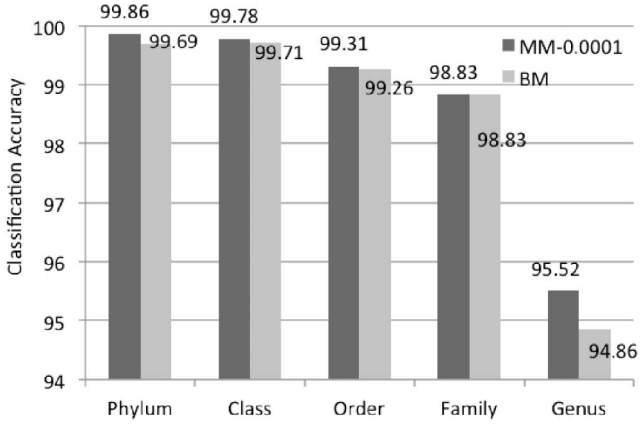

Fig. 1. The comparison of accuracies between the MM-0.0001 and the BM models on the full-length 16S rRNA gene sequence data set.

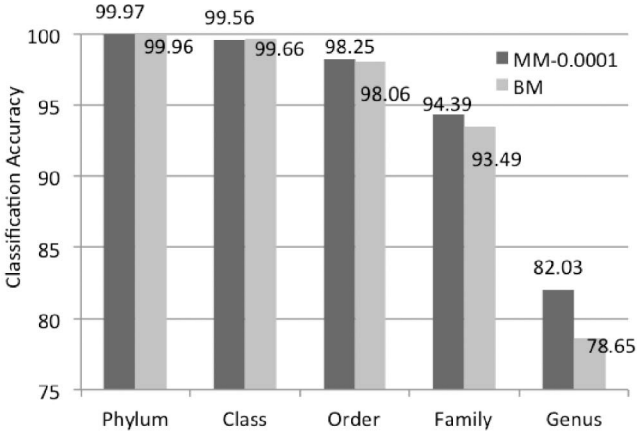

Fig. 2. The comparison of accuracies between the MM-0.0001 and the BM models on the full-length 28S LSU gene sequence data set.

gene sequence sets, respectively. The MM model outperforms the BM model in every taxon of the two gene sequence sets except for the rank family of the 16S rRNA and the rank class of the 28S fungal LSU. This finding suggests that if the flattening constant is properly chosen, the multinomial model will generally outperform the binomial model.

3.2 Short Read Fragment Comparison

Based on rRNA sequence alignment, the whole-region overlapping extraction method is utilized to obtain short gene sequence fragments for LOOCV testing: 250-bp is used to represent 454 Illumina reads, 400-bp is used to represent 454 Titanium reads, and 800-bp is used to represent Sanger reads. For each possible short read fragment length, the sequence extraction was carried out based on 25-base intervals that span the bp location relative to *Escherichia coli* for 16S rRNA and *S. cerevisiae* RDN 25-1 for 28S LSU rRNA as a reference sequence (without counting gap positions). An extracted sequence is represented by the position of its middle base. For instance, a 250-bp fragment that is extracted from base 401 to base 650 of a sequence will be specified by position 525. Extracted fragments were then used independently for LOOCV testing. Accordingly, the classification accuracy was tested across the entire 1,400-bp sequence length of fungal LSU rRNA and the entire 1,600-bp of 16S rRNA.

Fig. S1A in the supplemental material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2013.114, presents line charts of the accuracies of the BM, MM-0.01, MM-0.001, and MM-0.0001 models across the entire 16S rRNA gene sequence for the three considered fragment lengths. The 16S hypervariable regions V1 through V9 are shown just above the x-axis of Fig. S1A, available in the online supplemental material. When the fragment length is 250-bp, the multinomial model outperforms the binomial model for all extractions, except in the small area between positions 425 and 450 within hypervariable region V3. The MM-0.0001 model at position 550 yields the highest accuracy of 92.38 percent. The accuracies of the MM-0.001 and the MM-0.0001 models are generally not less than the accuracies of the other models for 250-bp fragments.

Fig. S1B, available in the online supplemental material, reveals that the accuracies for the 400-bp fragments ranges from 88 to 94 percent, and that the highest and the lowest accuracies are 93.76 percent achieved by the MM-0.001 model at position 275 and 88.66 percent achieved by the BM model at position 975, respectively. When the fragment length is 400-bp, every multinomial model considered in this study is more accurate than the binomial model. The predictive accuracy of the multinomial model typically increases as the flattening constant becomes smaller, and the curves for the MM-0.001 and MM-0.0001 models are very close.

When the extracted read fragment length is 800-bp, the accuracies range from 93 to 95.3 percent. The MM-0.001 model has the highest accuracy, 95.27 percent, at positions between 450 and 475. When compared to results of fragment lengths 250-bp and 400-bp, the accuracies of each model is relatively stable for fragment length of 800-bp. The main reason is that a 800-bp query read fragment is sufficiently long to provide the necessary information for classification. Fig. S1C, available in the online supplemental material, demonstrates that multinomial models outperform the binomial model at all positions.

Fig. S2, available in the online supplemental material, presents the accuracies of the BM, MM-0.01, MM-0.001, and MM-0.0001 models across the entire 28S LSU rRNA gene sequence for the three considered fragment lengths. Two fungal LSU rRNA hypervariable regions D1 and D2 are labeled on the x-axis of Figure S2A, available in the online supplemental material. The binomial model has the best performance before position 425 when the fragment length is 250-bp and before position 325 when the fragment length is 400-bp, as shown in Figs. S2A and S2B, available in the online supplemental material. Fig. S2C, available in the online supplemental material, shows that the binomial model no longer has a higher accuracy than various multinomial models in the leading positions. This result suggests that multinomial model is a better choice for long fragments. With reference to the y-axis of Fig. S2, available in the online supplemental material, the accuracies for short fragments on this gene sequence set are spread over a wide range. If the position of a short fragment is not located in the regions covered by the two hypervariables, and its length is not long enough, say less than 800-bp, then the probability of correct classification on this fragment read is likely to be relatively low.

When the length of a short read extracted from the D2 region is either 250-bp or 400-bp, the MM-0.0001 model achieves the highest accuracy. Since every extracted 800-bp short read before position 950 will include at least one of the hypervariables D1 and D2, the accuracy before position 950 is relatively stable, as shown in Fig. S2C, available in the online supplemental material. The improvement on the region after hypervariable D2 is larger because of the extremely low coverage over that region in the training set [13]. Overall, the MM-0.0001 model has the best performance for all three read fragment lengths.

4 Discussion

Feature extraction is essential for classifying gene sequence reads. Multinomial model counts the frequency of a k-mer word in a sequence read, while binomial model considers only whether a sequence read includes a k-mer word. Intuitively, a classifier can achieve a higher predictive accuracy when both training and testing phases employ the same model to calculate probability estimate for the naïve Bayesian classifier. Although the RDP classifier is computationally efficient and has a high predictive accuracy, it uses binomial model in the training phase but multinomial model in the testing phase. If the multinomial model can be used in the training phase, the estimate of $P(w_j^{(k)} \mid C_i)$ will be more reliable under which the computational efficiency of the naïve Bayesian classifier is preserved.

Two main factors can affect the predictive accuracy of the naïve Bayesian classifier: the model for estimating probabilities and the length of gene sequence reads. The multinomial model accompanied with longer length of gene sequence reads employs more thorough information carried in data for estimating probabilities, and hence, this combination should achieve the highest predictive accuracy. As shown in Table 3, the fact that this combination of the MM-0.0001 model with full-length read has the best performance in 16S SSU rRNA is, therefore, reasonable. In Fungal 28S LSU rRNA, since the predictive accuracy will be dramatically reduced by the sequence segment located after hypervariable D2, as can be seen from Fig. S2, available in the online supplemental material, the performance of the MM-0.0001 model with full-length read is not the best. When the length of a gene sequence read is not long enough, the probability estimates obtained from the multinomial model may not be reliable. In particular, the binomial model is very competitive when a short sequence read is outside the hypervariable regions.

The value of the flattening constant in the multinomial model can also affect the performance of the naïve Bayesian classifier. The purpose of the flattening constant is to ensure that every probability estimate is positive. The confidence level in the equality of the occurrence probabilities of all features is proportional to the flattening constant. Since the number of reads in a gene sequence set is much smaller than the number of possible features in this set, a large flattening constant will reduce the

TABLE 3
The Highest Accuracies Achieved by Various Models Applied on Various Lengths

	16S SSU rRNA				Fungal 28S LSU rRNA			
	250bp	400bp	800bp	Full length	250bp	400bp	800bp	Full length
BM	91.10%	92.85%	94.49%	94.86%	80.28%	80.79%	81.41%	78.65%
MM-0.01	91.59%	93.41%	94.99%	95.30%	80.26%	81.36%	82.38%	80.65%
MM-0.001	92.27%	93.77%	95.27%	95.51%	81.07%	82.04%	83.30%	81.80%
MM-0.0001	92.39%	93.66%	95.14%	**95.52%**	81.64%	82.54%	**83.74%**	82.04%

impact of the training reads. The relatively small predictive accuracies obtained by the MM-1 and MM-0.1 demonstrate that strong confidence about the noninformative assumption will have a negative impact on the performance of the naïve Bayesian classifier. When this flattening constant is sufficiently small, the resulting accuracy will be stable, because every prediction is primarily determined by the training data. This fact explains why the highest accuracies of the MM-0.01, MM-0.001, and MM-0.0001 models for a gene sequence set are quite close to each other, as shown in Table 3.

The multinomial model requires more time than does the binomial model to count the frequencies of k-mer words in the training data. However, this counting process can be completed by scanning a sequence set once. The additional effort required to use the multinomial model will, therefore, have only a tiny impact on the computational efficiency of the naïve Bayesian classifier.

5 Conclusion

Although the RDP classifier is efficient and has a competitive accuracy in classifying gene sequence reads, it employs the binomial model in the training phase but the multinomial model in the testing phases. Since the binomial and the multinomial models do not apply to exactly the same information that is embedded in the data, applying the same model in both phases would be more reasonable. This paper proposed the naïve Bayesian classifier that uses the multinomial model in both training and testing phases for calculating classification probabilities. A flattening constant for the multinomial model is introduced to ensure that every probability estimate will be positive for the naïve Bayesian classifier. Not only full-length sequence reads, but also short fragment reads with lengths 250-bp, 400-bp, and 800-bp are processed to evaluate the performance of the proposed classifier with respect to that of the RDP classifier. The experimental results conducted on 16S rRNA and 28S rRNA gene sequence sets show that the proposed classifier generally outperforms the RDP classifier when the flattening constant is small.

The number of instances for a specific class value in a sequence set can be small; for example, many class values in the genus level contain fewer than ten instances. In such a case, allowing different confidence levels on the features in a noninformative prior can be important role in classifying a sequence read. The prior-setting methods proposed by Wong [25], therefore, have the potentiality to improve performance of naïve Bayesian classifier for gene sequence data.

Acknowledgments

This research was supported by the National Science Council in Taiwan under Grant No. 99-2410-H-006-072-MY2.

References

[1] J.F. Araujo, A.P. de Castro, M.M. Costa, R.C. Togawa, G.J. Junior, B.F. Quirino, M.M. Bustamante, L. Williamson, J. Handelsman, and R.H. Kruger, "Characterization of Soil Bacterial Assemblies in Brazilian Savanna-Like Vegetation Reveals Acidobacteria Dominance," *Microbial Ecology,* vol. 64, pp. 760-770, May 2012.

[2] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D.R. Mende, G.R. Fernandes, J. Tap, T. Bruls, J.M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H.B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E.G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W.M. de Vos, S. Brunak, J. Dore, M. Antolin, F. Artiguenave, H.M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K.U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Merieux, R. MeloMinardi, C. M'Rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S.D. Ehrlich, and P. Bork, "Enterotypes of the Human Gut Microbiome," *Nature,* vol. 473, pp. 174-180, May 2011.

[3] N.A. Bokulich, C.M. Joseph, G. Allen, A.K. Benson, and D.A. Mills, "Next-Generation Sequencing Reveals Significant Bacterial Diversity of Botrytized Wine," *PLoS One,* vol. 7, article e36357, 2012.

[4] J.R. Cole, B. Chai, R.J. Farris, Q. Wang, S.A. Kulam, D.M. McGarrell, G.M. Garrity, and J.M. Tiedje, "The Ribosomal Database Project (RDP-II): Sequences and Tools for High-Throughput rRNA Analysis," *Nucleic Acids Research,* vol. 33, pp. D294-D296, Jan. 2005.

[5] J.R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, R.J. Farris, A.S. Kulam-Syed-Mohideen, D.M. McGarrell, T. Marsh, G.M. Garrity, and J.M. Tiedje, "The Ribosomal Database Project: Improved Alignments and New Tools for rRNA Analysis," *Nucleic Acids Research,* vol. 37, pp. D141-D145, Jan. 2009.

[6] E.F. DeLong, "Microbial Community Genomics in the Ocean," *Nature Rev. Microbiology,* vol. 3, pp. 459-469, June 2005.

[7] J. Dunbar, S.A. Eichorst, L.V. Gallegos-Graves, S. Silva, G. Xie, N.W. Hengartner, R.D. Evans, B.A. Hungate, R.B. Jackson, J.P. Megonigal, C.W. Schadt, R. Vilgalys, D.R. Zak, and C.R. Kuske, "Common Bacterial Responses in Six Ecosystems Exposed to 10 Years of Elevated Atmospheric Carbon Dioxide," *Environmental Microbiology,* vol. 14, pp. 1145-1158, May 2012.

[8] M. Egert, S. Marhan, B. Wagner, S. Scheu, and M.W. Friedrich, "Molecular Profiling of 16S rRNA Genes Reveals Diet-Related Differences of Microbial Communities in Soil, Gut, and Casts of Lumbricusterrestris L. (Oligochaeta: Lumbricidae)," *FEMS Microbiology Ecology,* vol. 48, pp. 187-197, May. 2004.

[9] S.A. Eichorst and C.R. Kuske, "Identification of Cellulose-Responsive Bacterial and Fungal Communities in Geographically and Edaphically Different Soils by Using Stable Isotope Probing," *Applied Environmental Microbiology,* vol. 78, pp. 2316-2327, Apr. 2012.

[10] P. Hugenholtz, B.M. Goebel, and N.R. Pace, "Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity," *J. Bacteriology,* vol. 180, pp. 4765-4774, Sept. 1998.

[11] A. Jumpponen, K.L. Jones, J. David Mattox, and C. Yaege, "Massively Parallel 454-Sequencing of Fungal Communities in *Quercus* spp. Ectomycorrhizas Indicates Seasonal Dynamics in Urban and Rural Sites," *Molecular Ecology,* vol. 19, no. Suppl 1, pp. 41-53, Mar. 2010.

[12] C.R. Kuske, C.M. Yeager, S. Johnson, L.O. Ticknor, and J. Belnap, "Response and Resilience of Soil Biocrust Bacterial Communities to Chronic Physical Disturbance in Arid Shrublands," *ISME J.,* vol. 6, pp. 886-897, Apr. 2012.

[13] K.L. Liu, A. Porras-Alfaro, C.R. Kuske, S.A. Eichorst, and G. Xie, "Accurate, Rapid Taxonomic Classification of Fungal Large-Subunit rRNA Genes," *Applied and Environmental Microbiology*, vol. 78, pp. 1523-1533, Mar. 2012.
[14] W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A.W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K.H. Schleifer, "ARB: A Software Environment for Sequence Data," *Nucleic Acids Research*, vol. 32, pp. 1363-1371, 2004.
[15] T.M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.
[16] H.E. O'Brien, J.L. Parrent, J.A. Jackson, J.M. Moncalvo, and R. Vilgalys, "Fungal Community Analysis by Large-Scale Sequencing of Environmental Samples," *Applied and Environmental Microbiology*, vol. 71, pp. 5544-5550, Sept. 2005.
[17] N.R. Pace, "A Molecular View of Microbial Diversity and the Biosphere," *Science*, vol. 276, pp. 734-740, May 1997.
[18] J. Ravel, P. Gajer, Z. Abdo, G.M. Schneider, S.S. Koenig, S.L. McCulle, S. Karlebach, R. Gorle, J. Russell, C.O. Tacket, R.M. Brotman, C.C. Davis, K. Ault, L. Peralta, and L.J. Forney, "Vaginal Microbiome of Reproductive-Age Women," *Proc. Nat'l Academy of Sciences USA*, vol. 108, no. Suppl 1, pp. 4680-4687, Mar. 2011.
[19] G.L. Rosen, E.R. Reichenberger, and A.M. Rosenfeld, "NBC: The Naive Bayes Classification Tool Webserver for Taxonomic Classification of Metagenomic Reads," *Bioinformatics*, vol. 27, pp. 127-129, Jan. 2011.
[20] M.L. Sogin, H.G. Morrison, J.A. Huber, D. Mark Welch, S.M. Huse, P.R. Neal, J.M. Arrieta, and G.J. Herndl, "Microbial Diversity in the Deep Sea and the Underexplored 'Rare Biosphere,'" *Proc. Nat'l Academy of Sciences USA*, vol. 103, pp. 12115-12120, Aug. 2006.
[21] P.J. Turnbaugh, M. Hamady, T. Yatsunenko, B.L. Cantarel, A. Duncan, R.E. Ley, M.L. Sogin, W.J. Jones, B.A. Roe, J.P. Affourtit, M. Egholm, B. Henrissat, A.C. Heath, R. Knight, and J.I. Gordon, "A Core Gut Microbiome in Obese and Lean Twins," *Nature*, vol. 457, pp. 480-484, Jan. 2009.
[22] Y. Van de Peer, S. Chapelle, and R. De Wachter, "A Quantitative Map of Nucleotide Substitution Rates in Bacterial rRNA," *Nucleic Acids Research*, vol. 24, pp. 3381-3391, Sept. 1996.
[23] Q. Wang, G.M. Garrity, J.M. Tiedje, and J.R. Cole, "Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy," *Applied and Environmental Microbiology*, vol. 73, pp. 5261-5267, Aug. 2007.
[24] C.R. Woese and G.E. Fox, "Phylogenetic Structure of the Prokaryotic Domain: the Primary Kingdoms," *Proc. Nat'l Academy of Sciences USA*, vol. 74, pp. 5088-5090, Nov. 1977.
[25] T.T. Wong, "Alternative Prior Assumptions for Improving the Performance of Naive Bayesian Classifiers," *Data Mining and Knowledge Discovery*, vol. 18, pp. 183-213, 2009.
[26] F. Yang, X. Zeng, K. Ning, K.L. Liu, C.C. Lo, W. Wang, J. Chen, D. Wang, R. Huang, X. Chang, P.S. Chain, G. Xie, J. Ling, and J. Xu, "Saliva Microbiomes Distinguish Caries-Active from Healthy Human Populations," *ISME J.*, vol. 6, pp. 1-10, Jan. 2012.