BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 4 2014, pages 449–456
doi:10.1093/bioinformatics/btt700

Sequence analysis

Advance Access publication December 24, 2013

Phylogeny-based classification of microbial communities

Olga Tanaseichuk[1,*], James Borneman[2] and Tao Jiang[1,3,*]

[1]Department of Computer Science and Engineering, [2]Department of Plant Pathology and Microbiology, University of California, Riverside, CA 92521 USA and [3]School of Information Science and Technology, Tsinghua University, Beijing 100084, China

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Next-generation sequencing coupled with metagenomics has led to the rapid growth of sequence databases and enabled a new branch of microbiology called comparative metagenomics. Comparative metagenomic analysis studies compositional patterns within and between different environments providing a deep insight into the structure and function of complex microbial communities. It is a fast growing field that requires the development of novel supervised learning techniques for addressing challenges associated with metagenomic data, e.g. sensitivity to the choice of sequence similarity cutoff used to define operational taxonomic units (OTUs), high dimensionality and sparsity of the data and so forth. On the other hand, the natural properties of microbial community data may provide useful information about the structure of the data. For example, similarity between species encoded by a phylogenetic tree captures the relationship between OTUs and may be useful for the analysis of complex microbial datasets where the diversity patterns comprise features at multiple taxonomic levels. Even though some of the challenges have been addressed by learning algorithms in the literature, none of the available methods take advantage of the inherent properties of metagenomic data.

Results: We proposed a novel supervised classification method for microbial community samples, where each sample is represented as a set of OTU frequencies, which takes advantage of the natural structure in microbial community data encoded by a phylogenetic tree. This model allows us to take advantage of environment-specific compositional patterns that may contain features at multiple granularity levels. Our method is based on the multinomial logistic regression model with a tree-guided penalty function. Additionally, we proposed a new simulation framework for generating 16S ribosomal RNA gene read counts that may be useful in comparative metagenomics research. Our experimental results on simulated and real data show that the phylogenetic information used in our method improves the classification accuracy.

Availability and implementation: http://www.cs.ucr.edu/~tanaseio/metaphyl.htm.

Contact: tanaseio@cs.ucr.edu or jiang@cs.ucr.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on September 10, 2013; revised on November 1, 2013; accepted on November 22, 2013

1 INTRODUCTION

The diversity of the microbial world had been hidden from the eyes of scientists until the advent of metagenomics. Despite the vital roles of microbes in our planet's ecology, evolution and human health, large populations of bacteria remain poorly characterized because the majority of bacterial species have not been successfully cultivated (Amann *et al*., 1995). The metagenomics approach has offered a remedy: bypassing the need for isolation and cultivation, all the sequences present in an environmental sample are sequenced simultaneously, making it possible to access the genetic information of otherwise hidden organisms. The field of metagenomics has further been advanced by the recent improvements in DNA sequencing technologies (MacLean *et al*., 2009) that allow millions of reads to be produced at a high speed and affordable costs. Deep sequencing allows for a high-resolution detection of rare species and provides an insight into phylogenetic composition and functional diversity of complex microbial populations with many low-abundance species. Along with the ability to access the genomes of rare species, next-generation sequencing technologies have also led to the rapid increase in the number and sizes of metagenomic sequencing projects. Exponential growth of sequence data has enabled comparative analysis of microbial communities, leading to a new branch of microbiology called comparative metagenomics.

Comparative analysis extends insights into the structure and function of microbial communities: it may help to identify community-specific properties of different environments as well as discriminative properties between different conditions, and to determine how microbial community composition is affected by specific environmental changes. Comparative metagenomics has broad implications for various fields of environmental science and human biology. It may help to address the intriguing question of identifiability of a core human microbiome (Turnbaugh *et al*., 2007), to understand the relationship between human microbiome and health and to study how microbial composition and function vary between distinct body sites and across the human population. Comparative analysis has already led to the discovery of three major enterotypes associated with the human gut microbiota (Arumugam *et al*., 2011) and the differences between lean and obese individuals (Turnbaugh *et al*., 2006).

Microbial communities can be compared at the levels of sequence composition, taxonomic diversity or functional potential. Taxonomic diversity provides detailed evolutionary information regarding the community composition and therefore has been a focus of many environmental and human studies (Costello *et al*., 2009; Lozupone and Knight, 2007). To access the taxonomic composition of a microbial community genomic

*To whom correspondence should be addressed.

sample, both single marker gene sequencing and whole community shotgun sequencing are widely used (Kembel *et al*., 2011). The 16S ribosomal RNA gene [or 16S ribosomal DNA (rDNA)] is a commonly used marker for bacterial identification due to its universal distribution among all bacterial species and a slow rate of sequence evolution. To reduce the dimensionality of large sequence datasets generated by high-throughput sequencing of 16S rDNAs, the reads are clustered into operational taxonomic units (OTUs) (Ye, 2011) that roughly represent taxa at phylogenetic levels defined by a user-defined sequence similarity cutoff. The abundance of each OTU is defined as the number of sequences in the OTU. Representative sequences from each OTU are chosen and used to assign taxonomy to the OTUs and to construct phylogenetic trees. Packages such as QIIME (Caporaso *et al*., 2010b) and Mothur (Schloss *et al*., 2009) provide integrated pipelines for the analysis described earlier in the text. The whole genome shotgun sequencing (i.e. metagenomics) approach can also be used to study microbial community composition. It offers a more global view of the community, but may not be deep enough to detect rare species in a sample, and is sensitive to the DNA extraction and sequencing protocols. Even though the two approaches may not always lead to the same conclusions about the community structure (Shah *et al*., 2011), they became standard tools in microbial community analysis. In this article, we will focus on samples based on the sequencing of 16S rDNAs although our proposed method can be easily extended to metagenomic samples based on whole genome sequencing.

For 16S rDNA datasets, each sample is represented as a list of OTUs and their frequencies. We will refer to these lists as *feature vectors*. A feature vector corresponds to one microbial community sample, and the vector's elements (*features*) characterize OTU frequencies in the sample. The evolutionary relationship between all of the OTUs is captured by a phylogenetic tree. The downstream analysis may involve the identification of compositional patterns across samples from similar environments as well as discriminatory features between different communities, associations between human bacterial communities and disease phenotypes, prediction of unknown labels for new samples and so forth. These tasks require the development of new supervised learning techniques that would take into consideration challenges associated with microbial community data. For example, features defined by OTUs do not necessarily represent specific taxonomic units because taxonomic levels are hard to define due to the fact that only relatively a small number of bacteria have been cultured. Moreover, it is hard to determine which taxonomic resolution level provides features with the best discriminative or predictive properties (Knights *et al*., 2011a). The environment-specific patterns may even comprise different lineages at varying phylogenetic depth. Finally, a low overlap in species between samples results in sparse and high-dimensional feature vectors.

The problem of classification of microbial communities is not well-studied, even though classification techniques have been widely used in the field of bioinformatics, including classification of microarray cancer samples (Glaab *et al*., 2010), gene expression profiles (Asyali *et al*., 2006), protein families (Yi *et al*., 2012) and so forth. Classification of metagenomic (or 16S rDNA) samples may have useful applications enabling efficient organization and search in rapidly growing metagenomic (or 16S rDNA) databases, detection of disease phenotypes in clinical samples and forensic identification.

One of the first applications of supervised learning techniques to comparative metagenomics (Yang *et al*., 2006) was the classification of soil and sediment samples according to environment types using support vector machines (SVM) and K-nearest neighbors algorithms. Recently, the feasibility of applying standard supervised classification techniques to metagenomic/16S rDNA data was studied on several benchmark datasets of human microbiota (Knights *et al*., 2011b). MetaDistance (Liu *et al*., 2011) is the first dedicated algorithm for multiclass classification of human microbiota. The algorithm combines the advantages of instance-based and model-based methods, such as K-nearest neighbors and SVM. The aforementioned methods proved to be efficient learning techniques and address some of the challenges associated with the properties of metagenomic/16S rDNA data. However, none of the methods have yet taken advantage of the inherent properties of metagenomic data, although phylogenetic information contained in metagenomic samples has proved to be useful in comparative metagenomics. For example, similarity measures that take into account phylogeny, e.g. UniFrac (Lozupone and Knight, 2005) and its generalized versions (Chang *et al*., 2011), outperform non-phylogenetic distance in their ability to recover natural clusters of microbial communities (Lozupone and Knight, 2008). To incorporate phylogeny into the similarity measure, UniFrac-based methods calculate the degree to which the input samples share branch length on a phylogenetic tree. Another example of a phylogeny-based similarity measure is the parsimony test (Schloss and Handelsman, 2006), which uses Fitch's parsimony algorithm to compute the number of minimal changes along the phylogenetic tree necessary to explain all the labels of the sequences. A recursive phylogenetic distance was defined in Meta-Storms (Su *et al*., 2012) for the purpose of fast indexing of metagenomic databases.

Although phylogeny provides important information about the natural hierarchical grouping of features, it has not yet been adopted in classification algorithms for microbial communities. To incorporate the underlying structural information among the features in learning problems, several regularization methods have been proposed. For example, in the problem of tumor class prediction from gene expression measurements, functional groups of genes form the natural structure of data and the combination of L_1 and L_2 norms were used to encode the groups and variables within the groups (Jacob *et al*., 2009). The group lasso penalty coupled with logistic regression (LR) was applied for classification problems with feature groupings and proved to be useful in short DNA motif modeling and splice site detection (Meier *et al*., 2008). Composite absolute penalty (Zhao *et al*., 2009) extended the grouping approach to deal with overlapping groups and with hierarchical orderings of the input variables that reflect the order in which the variables should be included in the solution. The idea of hierarchical grouping was later used for the multitask regression learning problem where a tree encodes relationships between the elements of a multidimensional-dependent variable, and a balanced weighting scheme to weight the hierarchically overlapping groups was proposed (Kim and Xing, 2010). Here, we show that the hierarchical grouping ideas (Kim and Xing, 2010; Zhao *et al*., 2009) can be efficiently applied

to the multiclass classification problem where information about the natural hierarchical grouping of features is available.

In this article, we proposed a new multiclass classification method for 16S rDNA sequence (or metagenomic) samples that takes advantage of the natural structure of microbial community data encoded by a phylogenetic tree. The leaf nodes represent features corresponding to individual OTUs, and the internal nodes may be considered as super features. The hierarchical structure captures similarities and differences among the features and allows for the consideration of features at different granularity levels, which may be desirable given that the features do not necessarily correspond to specific taxonomic units and only represent groups of similar sequences. Thereby, some environment-specific patterns may comprise features at multiple granularity levels. We proposed a multinomial LR model with a tree-guided penalty that incorporates the hierarchy in the feature space and presented an efficient optimization algorithm to learn the model parameters.

We applied our algorithm to several real datasets of 16S rDNA sequences from the human microbiota. We compared the classification performance of our method with several state-of-the-art learning algorithms, and showed that incorporating the natural structure of the microbial data results in a model with a better predictive power. We also performed a comprehensive analysis of the proposed method by applying it to simulated datasets of different complexity. Because the problem of classifying 16S rDNA sequence samples is relatively new, there are no simulated data generators that would generate the data appropriate for our goal. Current comparative studies use simulations where features are generated independently from each other rather than according to a phylogeny. Therefore, we introduced a new simulation approach that uses the phylogenetic structure of microbial communities to generate OTU count data and simulate scenarios where community-specific patterns may comprise features at multiple levels of granularity. We showed that by taking advantage of the phylogeny, our algorithm has a robust performance with respect to the choice of feature resolution, which corresponds to the selection of a similarity threshold when defining OTUs for the real data.

2 METHODS

We will first define a classical multinomial LR model associated with our classification problem for 16S rDNA sequence samples. Then, we will provide a background on how prior knowledge about the data, such as sparsity and natural grouping, can be effectively incorporated into learning models. Finally, we will present a model for the classification problem that takes advantage of the hierarchical groups of features and describe a fast optimization algorithm.

Consider a supervised learning problem where K microbial communities correspond to different environments or conditions. Each community (class) is represented by several samples. Let N be the total number of samples present in the dataset. A sample $x^i, i \in \{1, \ldots, N\}$ is characterized by feature values representing the abundance values of each OTU, $x^i = (x^i_1, x^i_2, \ldots, x^i_M)$, where M denotes the number of features. Each feature x^i_j of x^i is a continuous random variable. We will first consider features as independent variables, but will later add more definitions and extend the model to incorporate the information about the feature space encoded by a phylogenetic tree. Let y^i be a K-dimensional class variable, where each component takes on binary values, so that $y^i_j = 1$ if x^i belongs to community j, and $y^i_j = 0$ otherwise. Our goal is to model class labels as a function of the input features, i.e. to find a classification rule so that a new 16S rDNA sample x can be assigned to one of the classes. The goodness of the fit is measured by the loss function $L(x, y, \beta)$, where β is a vector of model coefficients. We formulate the problem as a multinomial LR so that the probability of a class label given the feature vector is modeled by the logistic function:

Algorithm 1: Coordinate descent algorithm.

begin
  for $k = 1, ..., K; j = 1, ..., M$ **do**
    $\beta_{kj} \leftarrow 0$;
    $\delta_{kj} \leftarrow 1$;
  for $t = 1, 2...$ **do**
    for $j = 1, ..., M$ **do**
      for $k = 1, ..., K$ **do**
        compute $\delta_{v_{kj}}$;
        $\delta_{\beta_{kj}} \leftarrow min(max(\delta_{v_{kj}}, -\delta_{kj}), \delta_{kj})$;
        $\beta_{kj} \leftarrow \beta_{kj} + \delta_{\beta_{kj}}$;
        $\delta_{kj} \leftarrow max(2|\delta_{\beta_{kj}}|, \delta_{kj}/2)$;

$$P(y_k = 1|x, \beta) = \frac{e^{x\beta_k^T}}{\sum_{k'} e^{x\beta_{k'}^T}},$$

where $\beta_k = (\beta^i_1, \beta^i_2, \ldots, \beta^i_M)$ is the vector of parameters corresponding to the k-th class.

A new sample is assigned to a class with the highest conditional probability estimate. The model coefficients β are obtained to better fit the observed data, i.e. to minimize the loss function defined as a log-likelihood:

$$L(x, y, \beta) = \sum_i \sum_k y^i_k x^i \beta_k - \sum_i \ln \sum_k e^{x^i \beta_k^T}.$$

To incorporate additional knowledge about the parameters, a Bayesian approach is often used. For example, the Gaussian prior favors parameter values that are close to 0, whereas the Laplace prior favors sparse solutions. Maximum a posteriori estimates of the model coefficients yield penalty terms in the form of the L_2 and L_1 norms for the Gaussian and Laplace priors, respectively. Finally, the optimal coefficients are obtained as a solution to a joint minimization of the loss function and the penalty function:

$$\arg\min_\beta \; L(x, y, \beta) + \lambda T(\beta),$$

where $T(\beta) = ||\beta||_2$ and $T(\beta) = ||\beta||_1$ for the two cases described earlier in the text.

In some situations, it may be desirable to consider some features as a group. Let β^{G_i} be a set of coefficients that correspond to the i-th group of features. The group penalty may be defined as a combination of norms for the groups and for coefficients within each group. To incorporate the hierarchical nature of the features, groups may be chosen to hierarchically overlap. Let T be a tree that reflects the hierarchical relationship among the features. Let us consider a node in a tree $v \in V$ and denote β^{G_v} to be a set of coefficients that correspond to features descendant from the node. As suggested in Kim and Xing (2010), the tree-guided penalty may be defined as follows:

$$T(\beta) = \sum_k \sum_{v \in V} \omega_v ||\beta^{G_v}_k||_2,$$

where the weight ω_v is associated with the node v and reflects a correlation within the group of features descending from v. Because each

coefficient β_j^i may correspond to multiple hierarchically overlapping groups of coefficients, it is important to penalize the coefficients equally. In particular, for each particular coefficient β_j^i, the weights of all the groups that contain this coefficient should sum to one. A weighting scheme that assures this property was proposed in Kim and Xing (2010):

$$\omega_\upsilon = \begin{cases} g_\upsilon \prod_{m \in Anc(\upsilon)} s_m & \text{if } \upsilon \text{ is an internal node,} \\ \prod_{m \in Anc(\upsilon)} s_m & \text{otherwise,} \end{cases}$$

Eventually, the solution to our problem is found as the minimum of the following function:

$$-\sum_i \sum_k y_k^i x^i \beta_k + \sum_i \ln \sum_k e^{x^i \beta_k^T} + \lambda \sum_k \sum_{\upsilon \in V} \omega_\upsilon ||\beta_k^{G_\upsilon}||_2.$$

Estimation of the model coefficients requires solving a convex optimization problem and therefore an efficient algorithm that scales up well to handle large high-dimensional data is necessary. Newton methods are intractable for high-dimensional data, whereas feature selection methods may not have good statistical foundations as most of them consider features in isolation, which is not desirable in our framework. Another challenge is that the penalty term is not differentiable at 0. A cyclic coordinate descent algorithm is a fast and simple algorithm that has been efficiently applied for the binary ridge LR model (Zhang and Oles, 2000) and extended to handle the non-smooth case of the lasso penalty (Madigan *et al.*, 2005). Our implementation for the tree-guided penalty is based on the aforementioned algorithms.

The cyclic coordinate descent method minimizes a function by cyclically updating each coefficient, while setting other coefficients fixed. The procedure continues until the convergence criterion is met. Each update involves a Newton step. Because Newton's method is based on the quadratic approximation of the objective function, large updates should be avoided when the quadratic approximation is poor. We use an updating rule suggested in the CLG algorithm (Zhang and Oles, 2000) and outlined in **Algorithm 1** (Madigan *et al.*, 2005). When computing the step size $\delta_{\upsilon_{kj}}$, unlike the CLG algorithm, we use the second derivatives directly rather then their upper bounds.

For a smooth function, the update $\delta_{\upsilon_{kj}}$ would be

$$\delta_{\upsilon_{kj}} = -\frac{\frac{\partial}{\partial \beta_{kj}}(L(x,y,\beta) + \lambda T(\beta))}{\frac{\partial^2}{\partial \beta_{kj}^2}(L(x,y,\beta) + \lambda T(\beta))}.$$

However, the penalty function does not have a derivative at $\beta_{kj} = 0$ because some of the terms in the penalty function contain the absolute value of β_{kj}. This happens because the L_2 norm $||\beta_k^{G_\upsilon}||_2$ degenerates to the L_1 norm $||\beta_{kj}||_1$ when either υ is a leaf node that corresponds to the j-th feature, or υ is an ancestral node of j and all the other coefficients descendant from υ are equal to 0. We will denote the set of nodes υ that have one of the aforementioned properties by V_{kj}. We follow the approach in Madigan *et al.* (2005) to handle the non-smooth penalty. Specifically, when $\beta_{kj} \neq 0$ we use the following update rule:

$$\delta_{\upsilon_{kj}} = -\frac{\frac{\partial}{\partial \beta_{kj}}(L(x,y,\beta) + \lambda D_1(k,j,\beta))}{\frac{\partial^2}{\partial \beta_{kj}^2}(L(x,y,\beta) + \lambda D_2(k,j,\beta))},$$

where

$$D_1(k,j,\beta) = \sum_{\substack{\upsilon \in Anc(j) \\ \upsilon \notin V_{kj}}} \omega_\upsilon \frac{\partial ||\beta_k^{G_\upsilon}||_2}{\partial \beta_{kj}} + \sum_{\upsilon \in V_{kj}} \omega_\upsilon Sign(\beta_{kj}),$$

$$D_2(k,j,\beta) = \sum_{\substack{\upsilon \in Anc(j) \\ \upsilon \notin V_{kj}}} \omega_\upsilon \frac{\partial^2 ||\beta_k^{G_\upsilon}||_2}{\partial \beta_{kj}^2}.$$

If updating β_{kj} results in a sign change, β_{kj} is set to 0. If $\beta_{kj} = 0$, we try updating β_{kj} in both directions, and if any of the updates results in a decrease in the objective function $L(x,y,\beta) + \lambda T(\beta)$, we update β_{kj} accordingly.

The implementation details that improve the computational efficiency include (i) maintaining a list of non-zero data entries for each dimension, (ii) storing the dot product of x^i and β_k and updating the result once β_k is updated, (iii) storing the sums of the exponentials $\sum_k e^{x^i \beta_k^T}$ and updating when the dot products in the exponents are updated, (iv) traversing the tree in the bottom-up manner to locate the sets V_{kj} of ancestors for each β_k^i to quickly find which terms in the penalty function degenerates to the L_1 norms and (5) computing partial sums of the hierarchically overlapping sets of coefficients in a top-down manner.

3 EXPERIMENTAL RESULTS

3.1 Synthetic framework and performance analysis

Simulated data are often used for extensive performance evaluation of algorithms under varying settings for which real data may not be available. Current data simulators that are used in comparative metagenomics simulate counts for each OTU independently not taking into consideration phylogenetic relationship between the species. For example, to evaluate the performance of MetaStats (White *et al.*, 2009), a statistical method for the comparison of clinical metagenomic samples (represented as counts of individual features such as organisms, genes and functional groups), read counts for each feature were generated according to negative binomial distributions independently from the other features. To evaluate phylogeny-based distance measures, such as UniFrac, microbial communities were represented as circles or ellipses (Chang *et al.*, 2011) with the overlap patterns between these ellipses expressing similarity between the communities. Species were simulated by sampling points from the interior of the ellipses. In this work, we propose a novel simulation approach to generate OTU count data that takes into account the input phylogeny and provides the flexibility for generating community-specific patterns at multiple granularity levels.

Our starting point is the common phylogenetic tree T that relates OTUs in all the 16S rDNA samples. For the ease of presentation, we will consider binary trees, although the model can be easily generalized to handle any trees. To generate samples for a class k, we traverse the tree systematically, deciding for each internal node υ what fraction of species would come from each of the subtrees rooted at the child nodes of υ. We associate two parameters with each node υ for each class k. Let μ_υ^k denote the average proportion of species that correspond to the subtree rooted at the left child node of υ in the k-th class, and let $(\sigma_\upsilon^k)^2$ denote the variance of this proportion within the class. A new class sample is generated by sampling the proportions of species at each node υ according to the normal distributions $N(\mu_\upsilon^k, (\sigma_\upsilon^k)^2)$. The parameters values μ_υ^k are in turn sampled from the normal distribution $N(\tilde{\mu}_\upsilon, \tilde{\sigma}_\upsilon^2)$, where parameters $\tilde{\sigma}_\upsilon^2$ characterizes the variance between the classes, and $\tilde{\mu}_\upsilon$ are some 'base' values that are initialized randomly. The simulation pipeline is outlined in Figure 1.

We control the within- and between-class variances using the parameters $(\sigma_\upsilon^k)^2$ and $\tilde{\sigma}_\upsilon^2$, respectively. We observed that in real-world datasets both variances tend to increase toward the leaves of the tree (see Supplementary Figs S1 and S2). To incorporate such a behavior into our framework, we define coefficients $\tilde{\gamma}$ and

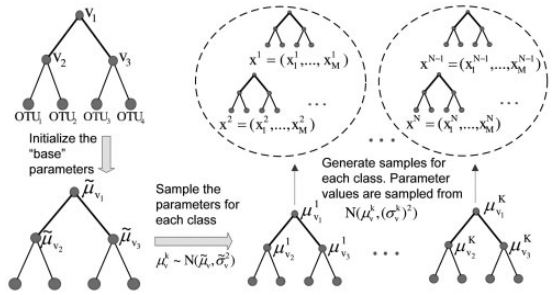

Fig. 1. The simulation pipeline

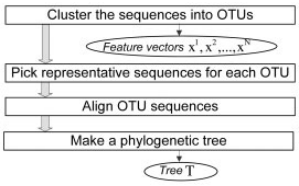

Fig. 2. The real data preprocessing pipeline. Rectangular boxes show the QIIME steps. Ellipses show the input data for our classification algorithm. First, all the reads are clustered into OTUs based on a user-defined similarity cutoffs using UClust (Edgar, 2010). For each sample, a feature vector of OTU frequencies is constructed. The most abundant sequence in each OTU is picked as the representative sequence. A multiple sequence alignment of the representative sequences is built using PyNAST (Caporaso *et al*., 2010a). Finally, the phylogenetic tree relating the OTUs is constructed from the multiple sequence alignment using FastTree (Price *et al*., 2010)

γ that describe the overall variances between and within the classes, respectively. We then sample the exact values of $\tilde{\sigma}_\upsilon$ and σ_υ^k at each tree node υ according to $N(0, \tilde{\gamma} d(\upsilon))$ and $N(0, \gamma d(\upsilon))$, where $d(\upsilon)$ is the distance between υ and the tree root. The parameters $\tilde{\gamma}$ and γ influence the difficulty of the classification problem, which is proportional to γ and inversely proportional to $\tilde{\gamma}$.

3.2 Comparisons

We compared the performance of our proposed method with some of the popular classification techniques such as SVMs, LR and Random Forests (RFs), and with a dedicated classification algorithm for metagenomic (or 16S rDNA sequence) samples, MetaDistance. SVM (Ben-Hur *et al*., 2008) is a robust and powerful classification method that has widespread applications in many fields including computational biology. The idea behind the SVM technique is to find the separating hyperplane in a feature space with the largest margin between the classes. We considered the L_1-regularized version of SVMs due to its ability to handle high-dimensional sparse data. RF (Boulesteix *et al*., 2012) is another popular classification algorithm that consists of a collection of tree predictors and makes the overall prediction by the majority voting. The algorithm is well known for its capability of dealing with small sample sizes, high-dimensional feature space and complex data. LR (Wu *et al*., 2009) has been widely used in statistics for many years and has recently received much attention in the machine learning community. In particular, its L_1-regularized version is known to have good generalization performance in the presence of many irrelevant features. MetaDistance (Liu *et al*., 2011) is a classification method for samples of 16S rDNA sequence counts. It is based on simultaneously minimizing the intraclass distance and maximizing the interclass distance by combining instance-based and model-based learning techniques. For the comparison with the SVM and LR classifiers, we used the implementations in the Machine Learning Python (MLPY) Python package (Albanese *et al*., 2012), for the RF classifier, we used scikit-learn Python package (Pedregosa *et al*., 2011) and for MetaDistance, we used the Matlab implementation (Liu *et al*., 2011). For all the classifiers, we performed 10-fold cross-validation to find the optimal model parameters (see Supplementary Table S1) and 3-fold (or 5-fold) cross-validations for the performance evaluation on simulated (or real, respectively) datasets.

3.3 Performance on simulated data

We simulated datasets of 2, 3, 5 and 10 classes with 20 samples for each class. To generate datasets of different complexity, we

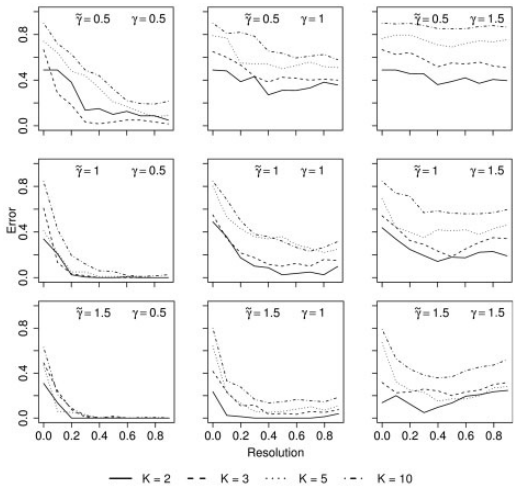

Fig. 3. Performance for varying number of classes and within- and between-class variances

considered the variance coefficients $\tilde{\gamma}$ and γ of 0.5, 1 and 1.5. A complete binary tree of height 10 was used to guide the simulation. For each parameter set, we generated datasets with different granularity by cutting the tree at different heights to produce feature vectors with different resolutions. The performance of our algorithm for all the parameter sets and resolution levels is shown on Figure 3. We observed that the classification task became harder when the number of classes increased. Also, as expected, the classification error rate reduced when γ was decreased and $\tilde{\gamma}$ was increased. We also observed that, for some parameter settings, the best performance was achieved at the highest resolution levels, whereas for the other settings the best performance might be achieved at some intermediate resolutions. This happened when the between-class variance was dominated by the within-class variance at high resolution levels, leading to an increased overlap between the classes. The use of the information about the hierarchical grouping of the features makes it possible to take advantage of the lower resolution feature space where the classes may be better separated. Figure 3 shows that the accuracy of our method at the highest

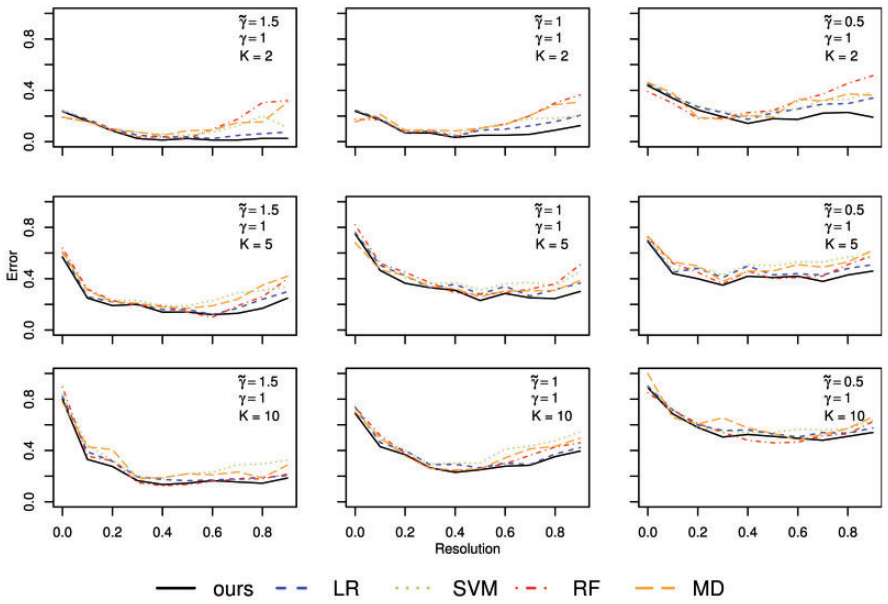

Fig. 4. Comparison with LR, SVM, RF and MetaDistance on simulated datasets for varying number of classes and between-class variances. The top, middle and bottom plots correspond to datasets with 2, 5 and 10 classes, respectively. The within-class variance $\gamma = 1$. The between-class variance $\tilde{\gamma}$ is 1.5, 1 and 0.5 on the left, middle and right plots

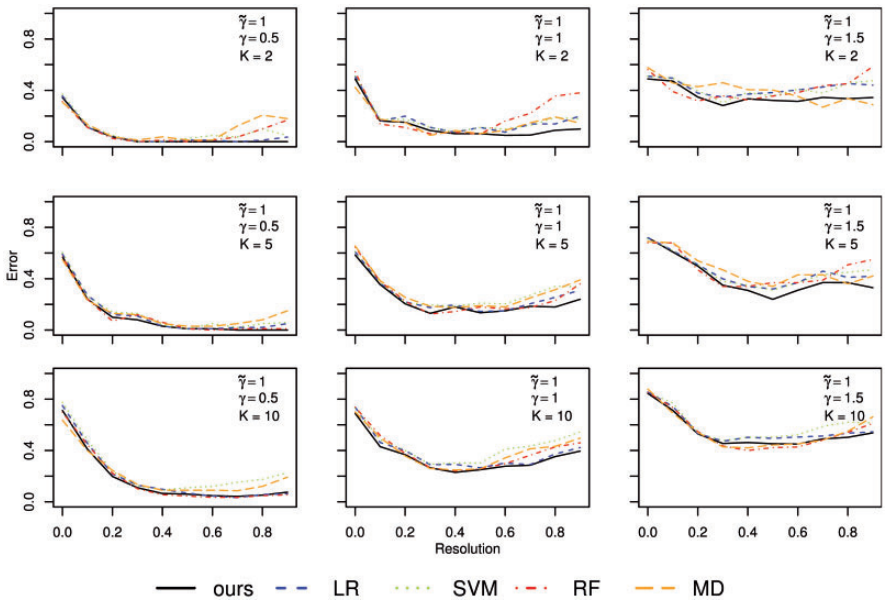

Fig. 5. Comparison with LR, SVM, RF and MetaDistance on simulated datasets for varying number of classes and within-class variances. The top, middle and bottom plots correspond to datasets with 2, 5 and 10 classes, respectively. The between-class variance $\tilde{\gamma} = 1$. The within-class variance γ is 0.5, 1 and 1.5 on the left, middle and right plots

resolution level was reasonable compared with the accuracy at the optimal resolution level.

We compared our algorithm with the other classifiers under a variety of parameter sets and plotted the results in Figures 4 and 5. The mean and standard deviations of the relative performance of the algorithms at the highest resolution level are shown in Table 1. Overall, LR performed reasonably well in most of the test cases, whereas SVM had an inferior performance compared with the other algorithms. RF performed poor when the number of classes was small, but improved significantly when the number of classes increased. In most of the test cases, our algorithm achieved the best performance and was more robust with respect to the resolution level. That is, when the optimal classification accuracy was achieved at some intermediate resolution, the performance at the highest resolution was still reasonably close to the optimal. Figures 4 and 5 show that the performance

Table 1. Relative performance of our algorithm compared with LR, SVM, RF and MetaDistance across 30 simulated datasets for $\tilde{\gamma} = 1$, $\gamma = 1$ and $K = 5$

LR		SVM		RF		MD	
Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
−4.3	3.9	−12.3	3.7	−10.0	5.8	−5.0	3.3

difference between the algorithms is especially noticeable at the highest resolution.

3.4 Performance on real data

We used three real datasets of the human microbiota to evaluate our algorithm. All the datasets were taken from the 16S ribosomal RNA sequencing studies and preprocessed as shown in Figure 2 using the QIIME software (Caporaso *et al*., 2010b). Using different OTU similarity cutoffs, we generated feature vectors of different granularity for each of the datasets. The dataset *D*1 is described in (Costello *et al*., 2009) and comprises samples from six major body areas: external ear, gut, hair, nostril, oral cavity and skin. The second dataset *D*2 contains gut samples from lean, obese and overweight subjects (Turnbaugh *et al*., 2009). The third dataset *D*3 is described in the study of microbiota in healthy adults (Human Microbiome Project Consortium, 2012) and contains samples from five body habitats: oral, gastrointestinal, urogenital, nasal and skin. Datasets *D*1 and *D*3 represent relatively easy classification problems because they comprise microbial communities from different body sites, which are known to be significantly different. On the other hand, *D*2 illustrates an example of a more challenging classification task because the classes correspond to microbial communities from the same body habitat and thus are similar. To support the conclusions about the difficulty of the problems represented by the three datasets, we have computed the average within- and between-class variances for each dataset (see Supplementary Figs S1 and S2). We observe that, compared with *D*1 and *D*3, dataset *D*2 on average has a lower between-class variance, which confirms that the corresponding classification problem is more challenging.

The classification error rates for all the classification algorithms at all the granularity levels are shown in Table 2. On average, all the algorithms show a better performance at higher resolution levels that correspond to the higher OTU similarity cutoffs. Our algorithm shows a comparable performance at the lower resolution levels and outperforms the other methods at the higher resolution levels.

We did not evaluate the computational efficiency of the algorithms systematically because they are implemented in different programming languages. Some of the entries are missing in Table 2 due to overly long processing time. More specifically, we were unable to preprocess the dataset *D*3 at 97% similarity cutoff with QIIME on our desktop PC. Therefore, we did not run any of the classification algorithms on the dataset at 97% similarity cutoff. MetaDistance was two orders of magnitudes slower than the other algorithms. For example, the training step for the dataset *D*1 with 90% OTU similarity cutoff for just one set of parameters took ∼1 s for LR and SVM, 3 s for our algorithm, 5 s for RF and 12 min for MetaDistance. Therefore, running cross-validation on multiple sets of parameters would take days or weeks to compute on some datasets for MetaDistance.

Table 2. Comparison of the error rates (%) with LR, SVM, RF and MetaDistance classifiers on real datasets

Dataset	Alg.	65	70	75	80	85	90	95	97
D1	Ours	12.9	**10.4**	**8.2**	**7.1**	**6.7**	**5.7**	**5.5**	**5.3**
	SVM	12.9	11.0	10	8.8	7.0	6.9	6.5	6.3
	LR	12.9	10.8	9.6	8.2	7.6	7.1	6.5	6.1
	RF	**12.7**	**10.4**	10.0	9.8	9.4	8.2	7.8	7.6
	MD	14.9	12.2	10.8	8	6.9	6.7		
D2	Ours	29.2	**27.4**	**24.6**	**24.6**	23.9	**21.0**	**19.2**	**16.7**
	SVM	30.3	29.6	26.7	28.9	26.7	26.4	26.7	23.1
	LR	29.9	29.2	27.7	28.5	25.6	26.0	23.9	21.4
	RF	**28.5**	29.2	26.0	25.6	24.5	26.3	26.7	27.8
	MD	30.2	30.2	30.2	26.0	**23.1**	22.1	23.4	20.9
D3	Ours	9.3	10.4	**7.2**	**4.9**	**3.8**	**3.5**	**3.3**	
	SVM	8.7	10.5	**7.2**	5.7	4.4	4.1	4.5	
	LR	**8.3**	10.2	7.7	5.4	4.3	4.5	3.8	
	RF	8.7	**10.0**	7.4	5.7	4.9	5.1	5.1	
	MD	12.8	10.3	8.7	6.2				

Note: The bold values indicate the best performance among all five methods.

4 CONCLUSION

We proposed a new classification method for 16S rDNA sequence samples that uses the natural structure of microbial community data encoded by a phylogenetic tree. We showed that using the phylogenetic information leads to an improved classification accuracy compared with the state-of-the-art classification algorithms. Unlike many popular classification methods, which consider features (or OTU frequencies) in isolation, our method takes advantage of the similarities between OTUs encoded by the phylogenetic tree. We applied the algorithm to classify samples obtained from 16S rDNA studies, but the approach can also be used to classify metagenomic samples obtained by whole genome sequencing. The algorithm only requires frequencies of taxonomic groups in each sample and a phylogenetic tree that relates these groups.

Even though our model does take advantage of the natural hierarchical grouping of OTUs, features are still defined as individual OTUs (the leaves of the phylogenetic tree). Nevertheless, a microbial community may be characterized by several lineages of varying phylogenetic depth. Knights *et al*. (2011a) gives an example of a hypothetical disease characterized by concurrent over-representation of a particular phylum, genus and species. Therefore, rather than defining features as tree leaves, a subset of intermediate nodes of the phylogenetic tree may lead to more powerful classification models for complex microbial

communities. Modeling environment-specific features at multiple granularity levels explicitly is a potential direction for future work.

ACKNOWLEDGEMENT

The authors are grateful to the anonymous referees for their thorough reviews and constructive comments.

Funding: National Science Foundation (grant DBI-1262107) (in part) and a UCR Seed grant.

Conflict of interest: none declared.

REFERENCES

Albanese,D. *et al.* (2012) MLPY: machine learning python. arXiv:1202.6548v2.

Amann,R.I. *et al.* (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, **59**, 143–169.

Arumugam,M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.

Asyali,M.H. *et al.* (2006) Gene expression profile classification: a review. *Curr. Bioinform.*, **1**, 55–73.

Ben-Hur,A. *et al.* (2008) Support vector machines and kernels for computational biology. *PLoS Comput. Biol.*, **4**, e1000173.

Boulesteix,A.-L. *et al.* (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, **2**, 493–507.

Caporaso,J.G. *et al.* (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, **26**, 266–267.

Caporaso,J.G. *et al.* (2010b) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Chang,Q. *et al.* (2011) Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, **12**, 118.

Human Microbiome Project Consortium. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Costello,E.K. *et al.* (2009) Bacterial community variation in human body habitats across space and time. *Science*, **326**, 1694–1697.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Glaab,E. *et al.* (2010) Learning pathway-based decision rules to classify microarray cancer samples. In: *German Conference on Bioinformatics 2010, of Lecture Notes in Informatics*. Vol. 173, Gesellschaft fuer Informatik, Germany, pp. 123–134.

Jacob,L. *et al.* (2009) Group lasso with overlap and graph lasso. In: *ICML'09: Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, New York, NY, pp. 433–440.

Kembel,S.W. *et al.* (2011) The phylogenetic diversity of metagenomes. *PLoS One*, **6**, e23214.

Kim,S. and Xing,E.P. (2010) Tree-guided group lasso for multi-task regression with structured sparsity. In: *Proceedings of the 27th International Conference on Machine Learning*. Haifa, Israel.

Knights,D. *et al.* (2011a) Human-associated microbial signatures: examining their predictive value. *Cell Host Microbe*, **10**, 292–296.

Knights,D. *et al.* (2011b) Supervised classification of human microbiota. *FEMS Microbiol. Rev.*, **35**, 343–359.

Liu,Z. *et al.* (2011) Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Bioinformatics*, **27**, 3242–3249.

Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.

Lozupone,C.A. and Knight,R. (2007) Global patterns in bacterial diversity. *Proc. Natl Acad. Sci. USA*, **104**, 11436–11440.

Lozupone,C.A. and Knight,R. (2008) Species divergence and the measurement of microbial diversity. *FEMS Microbiol. Rev.*, **32**, 557–578.

MacLean,D. *et al.* (2009) Application of'next-generation' sequencing technologies to microbial genetics. *Nat. Rev. Microbiol*, **7**, 287–296.

Madigan,D. *et al.* (2005) Bayesian multinomial logistic regression for author identification. In: *Maxent Conference*. pp. 509–516.

Meier,L. *et al.* (2008) The group lasso for logistic regression. *J. R. Stat. Soc. B Stat. Methodol.*, **70**, 53–71.

Pedregosa,F. *et al.* (2011) Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Price,M.N. *et al.* (2010) FastTree 2 Approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Schloss,P.D. and Handelsman,J. (2006) Introducing TreeClimber, a test to compare microbial community structures. *Appl. Environ. Microbiol.*, **72**, 2379–2384.

Shah,N. *et al.* (2011) Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. In: *Proceedings of the Pacific Symposium on Biocomputing*. Hawaii, Kohala Coast, pp. 165–176.

Su,X. *et al.* (2012) Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. *Bioinformatics*, **28**, 2493–2501.

Turnbaugh,P.J. *et al.* (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.

Turnbaugh,P.J. *et al.* (2007) The Human Microbiome Project. *Nature*, **449**, 804–810.

Turnbaugh,P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

White,J.R.R. *et al.* (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.*, **5**, e1000352.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Yang,C. *et al.* (2006) An ecoinformatics tool for microbial community studies: supervised classification of amplicon length heterogeneity (ALH) profiles of 16S rRNA. *J. Microbiol. Methods*, **65**, 49–62.

Ye,Y. (2011) Identification and quantification of abundant species from pyrosequences of 16S rRNA by consensus alignment. *Proc. (IEEE Int. Conf. Bioinformatics Biomed.)*, **2010**, 153–157.

Yi,G. *et al.* (2012) Supervised protein family classification and new family construction. *J. Comput. Biol.*, **19**, 957–967.

Zhang,T. and Oles,F.J. (2000) Text categorization based on regularized linear classification methods. *Inf. Retr.*, **4**, 5–31.

Zhao,P. *et al.* (2009) The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Stat.*, **37**, 3468–3497.