Supervised Learning to Predict Geographic Origin of Human Metagenomic Samples

Christopher Malow CMALOW@STANFORD.EDU

Abstract

Metagenomic studies of human fecal samples have used whole genome shotgun sequencing and 16S ribosomal RNA (rRNA) sequencing to characterize the composition of gut microbiota among geographically disparate populations, highlighting differences in the relative abundance of specific classes of microbes associated with diet, cultural factors, and environmental exposure. In this project, we investigate supervised machine learning techniques, trained on a subset of human gut microbiome sequencing data from three distinct geographic regions, to develop a prediction model that attempts to classify test samples from the same populations based on their geographic origin. We show that differences in the relative abundance of microbial populations among the samples can be used to construct multiclass prediction models that differentiate among samples from three geographic populations with approximately 90% accuracy.

As predicted, many of the learned parameters of the classifiers reflect previously documented associations between human populations and dietary and geographic factors. As reference samples for a wider range of geographic regions become available, similar classifiers could be used as forensics tools to link people, animals, or objects together with geographic locations, generating investigative leads for law enforcement, or as a tool to characterize migration patterns and the impact of the environment on human communities.

1. Background

The large, diverse population of microbes within the human gut has been associated with human metabolic capabilities, resistance to pathogens, and gastrointestinal development (Backhed et al., 2005). Metagenomic studies of human fecal samples have used whole genome shotgun sequencing and 16S ribosomal RNA (rRNA) sequencing to characterize the composition of gut microbiota among geographically disparate populations, highlighting differences in the relative abundance of specific classes of microbes associated with diet, cultural factors, and environmental exposure (Qin et al., 2010; Yatsunenko et al., 2012).

As samples from a wider range of geographic locations become available, supervised machine learning models indicative of geographic patterns in the human gut microbiome can be used as forensics tools to link people, animals, or objects together with geographic locations (Gunn and Pitt, 2012), or as tools to characterize migration patterns and the impact of the environment on human communities (Parks et al., 2013). Although the ability of the human microbiome to change in response to environmental conditions results in variability over time, it is precisely because of this variability in response to environmental factors that methods based on microbiota can complement established DNA-based forensic methods for identification; analysis of microbiome data may provide clues as to a person's recent whereabouts and interactions with other people and objects and provide investigative leads for law enforcement (Gunn and Pitt, 2012).

In this project, we use supervised machine learning techniques, trained on a subset of human gut microbiome sequencing data from three distinct geographic regions, to develop prediction models that attempt to classify test samples from the same populations based on their geographic origin. We hypothesized that differences in the presence or relative abundance of microbial populations among the samples would inform a model that could predict the geographic origin of human microbiome samples drawn from these three geographic regions, and that the resulting parameters

would reflect findings from previous studies on the role of particular classes of bacteria in human metabolism across different dietary and geographic conditions.

2. Materials and Methods

2.1. Data Collection and Processing

Recent studies (Yatsunenko et al., 2012) have collected and processed 16S rRNA sequence data from fecal samples (one sample per individual) drawn from 528 individuals residing in the Amazonas region of Venezuela, rural Malawian communities in Africa, and inhabitants of US metropolitan areas, including St. Louis, Philadelpha, and Boulder. The previously collected dataset, hosted online in the MG-RAST metagenomic analysis database (Meyer et al., 2008), characterizes the taxonomic/phylogenetic composition of microbiota in each sample using the sequence data as taxonomic markers.

Initial data analysis and retrieval was performed using the online MG-RAST analysis toolkit, which allows abundance data to be computed at any available taxonomic level for samples in the collection. An index of sample metadata and collections of abundance data at the genus level for the 528 individuals/samples in the study was downloaded from the MG-RAST repository. Using the R programming language, data provided on the abundance of each genus within each sample was used to construct a set of features corresponding to the relative abundance of each genus within each sample. Hits within "unclassified" or "unassigned" categories, representing a small portion of unknown genera identified in each sample, were excluded from the feature set using regular expressions. The relative abundance of a genus was calculated by dividing the number of hits for the genus by the total hits for the sample. After processing in this manner, the dataset included 959 potential features for each sample, with each feature corresponding to the relative abundance of a microbial genus detected in at least one of the samples.

2.2. Analysis Methods

This project explored four different supervised machine learning methods as candidate prediction models. Models for Naive Bayes and Support Vector Machine (SVM) classifiers were constructed using the e1071 package in R (Meyer et al., 2012), logistic regression models were constructed with the generalized linear model (glm) function within R, and Bayesian logistic regression models were constructed using the arm package in R (Gelman et al., 2013). The cost parameter and kernel selection for the multiclass SVM classifier were optimized against the 10-fold cross-validation error of the multiclass model (Figure 3); the final SVM model was trained using a linear kernel and normalized with cost factor $C = 1$, but radial, polynomial, and sigmoid kernels were also evaluated during the optimization process. The Naive Bayes model was trained with Laplace smoothing ($\alpha = 1$) enabled. Multiclass SVM and Bayesian logistic regression classifiers were constructed using the one-versus-one method, included in the e1071 package SVM classifier and implemented in R for the Bayesian logistic regression classifier.

2.3. Cross-Validation and Feature Subsets

To evaluate candidate machine learning algorithms (Table 1 and Model Selection, below), samples corresponding to communities in Malawi were excluded to permit binary classification between the US and Venezuelan samples, and a randomly selected sample consisting of 30% of the subset of samples was withheld and used to determine the reported testing error. Of the 959 potential features for each sample, subsets of smaller numbers of features were selected beginning with features that had the largest average relative abundance across the samples; thus the 20-feature subset corresponded to the 20 features with the largest average relative abundance.

Models selected for further investigation (Table 2 and Multiclass Model Optimization, below) incorporated data from all three regions (Malawi, US, and Venezuela) and used 10-fold cross-validation for reported accuracies; SVM cross-validation was performed with the built-in functionality of e1071 package, and cross-validation for the Bayesian logistic regression model was implemented in R using the cvTools package for selection of cross-validation folds.

3. Results and Discussion

3.1. Model Selection

In our preliminary analysis, excluding samples corresponding to Malawi left a training set of 290 samples and a test set of 124 samples. With 959 potential features and only 290 available training examples, the feature set used in the model played a large role in the observed accuracy of several of the models, so each model was tested against feature sets consisting of 20, 60, 400, and 959 microbial genera (Table 1).

After initial comparison of the Naive Bayes, Logistic GLM, and SVM models showed a significant dropoff in performance of the Naive Bayes and logistic regression models with more complex feature sets, most likely due to high variance, the Bayesian logistic regression model

Table 1. Initial Evaluation of Candidate Algorithms: Classification accuracies for various models on the US/Venezuela dataset are reported. Bayesian logistic regression had the highest overall accuracy (94.4%) on the holdout dataset. Logistic regression models marked with * did not converge.

Feature Set Size:	20	60	400	959
Naive Bayes	85.4	91.1	89.5	35.5
Logistic GLM	91.9	88.7*	57.3*	48.4*
Bayesian Logistic	89.5	88.7	94.4	94.4
SVM-Linear	87.9	87.9	87.9	87.9

was added to explore whether a maximum a posteriori (MAP) estimate model could incorporate more of the information contained in the low-abundance features without overfitting.

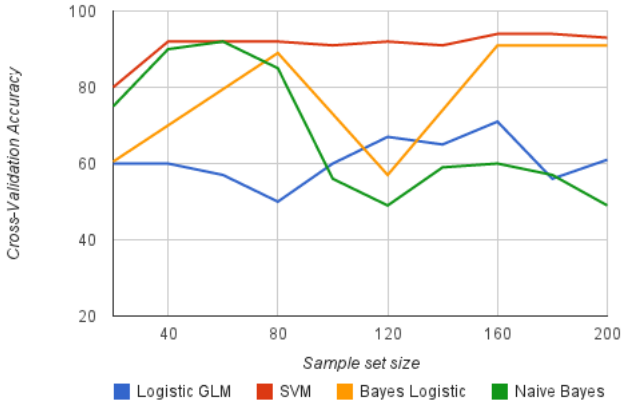

Figure 1. Learning curves for binary classifiers on the US/Venezuela dataset using the full set of 959 candidate features. The standard logistic regression and Naive Bayes models showed a dropoff or lack of improvment in performance with larger sample sizes, suggesting a high-variance fit was preventing improvments in accuracy.

The relatively high accuracies observed for all four models (with various feature set sizes) indicates accurate binary classifiers for the geographic origin of samples drawn from the two regions can be constructed based on 16S rRNA data. Plotting learning curves for select models, accuracies continued to improve as the sample size approached 150, suggesting that, in general, geographic microbiome classifiers based on small sample sizes may result in reduced accuracy (Figures 1 and 2). The strong performance of the Bayesian logistic regression model suggests that MAP estimate models can provide improvements in accuracy by incorporating information from the lower-abundance features.

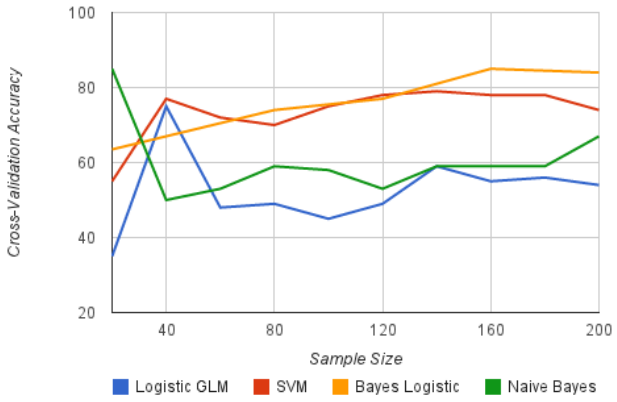

Figure 2. Learning curves for binary classifiers on the Malawi/Venezuela dataset using the full set of 959 candidate features. Bayesian logistic and SVM classifiers performed the best overall, with most of the imporovements in accuracy occurring before the sample size reached 130.

3.2. Multiclass Model Optimization

The Bayesian logistic regression model was selected for further investigation based on its high accuracy and the potential for its weight parameters to be interpreted with respect to known functional and geographic associations of specific bacterial genera (see Functional Analysis, below); the SVM model was also selected for further investigation for its short execution time and potential to be used for larger, more complex datasets in the future.

10-fold cross-validation accuracies for both the Bayesian logistic regression and SVM classifiers are given in Table 2. Lower accuracies were observed for the Malawi/Venezuela binary classifier than for the Malawi/US and US/Venezuela classifiers (Table 3), most likely due to greater similarities in the diets of individuals in Malawi and Venezuela (Yatsunenko, et al., 2012). With an accuracy of 88.8% using the Bayesian logistic classifier and 90.3% using the linear SVM model, the results suggest that supervised machine learning classifiers can discriminate between the geographic origin of samples drawn from the three regions with acceptable accuracy, although performance is more modest when discriminating between two groups with similar diets.

3.3. Functional Analysis

One potentially interesting outcome of constructing prediction models with Bayesian logistic regression is the ability to interpret the weight parameters in

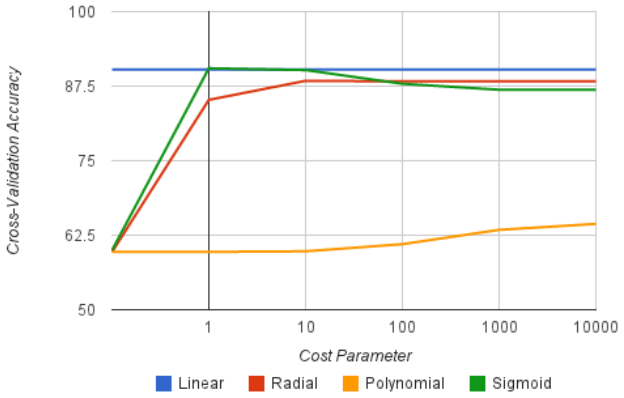

Figure 3. Kernel selection and cost parameter optimization for the multiclass (Malawi/US/Venezuela) SVM classifier. The linear SVM model showed high accuracy for all tested values of the cost parameter.

Table 2. Final Results: 10-fold cross-validation accuracies for binary models on pairwise subsets and multiclass models on the full Malawi/US/Venezuela dataset. The lowest accuracies were observed for the Malawi/Venezuela subset, most likely due to similar high-fiber, primarily vegetarian diets.

Dataset	Bayes Logistic	SVM
Malawi/US	93.5	96.5
US/Venezuela	94.4	96.6
Malawi/Venezuela	84.5	87.3
Malawi/US/Venezuela	88.8	90.3

light of previous domain-specific observations about the prevalence of microbial genera in different human populations (Table 3). For example, gut microbiota samples with high *bacteroides* abundance have been linked to diets high in animal protein and saturated fat, characteristic of US and other Western diets, while samples with high *prevotella* abundance are associated with diets high in carbohydrates, fiber, and simple sugars (Wu et al., 2011).

The large weights of parameters associated with *roseburia* and *butyrivibrio* in the Malawi/Venezuela model suggests that the model may be discriminating between the two populations using more subtle discriminating factors than protein and saturated fat composition in the diet. Recent microbiome studies of urban and rural communities in Russia identified *roseburia* as a key discriminating factor among rural populations in Tatarstan, Tyva, and Omsk, for instance (Tyakht et al., 2013).

Table 3. Confusion matrix of classification errors for the multiclass SVM model based on 10-fold cross validation of the available dataset. Errors were most frequent when differentiating between the Malawi and Venezuela samples.

	Actual		
Prediction	Malawi	US	Venezuela
Malawi	-	5	16
US	6	-	11
Venezuela	17	14	-

4. Conclusion

Genus-level abundance data from 16S rRNA sequencing of the geographically separated samples allows supervised machine learning methods using Bayesian logistic regression and SVM to classify samples with 88-90% accuracy. The results indicate that supervised machine learning models that control for variance, such as MAP estimate algorithms and SVMs, can provide superior accuracy for microbiome geographic classification tasks compared to maximum likelihood methods, based on their ability to incorporate information from a larger feature set without overfitting.

Several features of the models point to opportunities for future improvements. First, the models are constrained by the availability of reference samples drawn from people in geographic regions of interest. Although geographically-labeled samples are only available for specific regions, several recent studies have expanded the available set of collections by sharing data on Korean (Nam et al., 2011), Russian (Tyakht et al., 2013), and Chinese (Ling et al., 2013) individuals. Future work could incorporate data from additional geographic groupings as they become available.

Second, although classifiers with accuracies greater than 88% can provide useful information on the presumptive geographic origin of microbiome samples, higher accuracies may allow for a broader range of applications in scientific and forensic work. Since data are available at multiple taxonomic levels, machine learning algorithms that incorporate hierarchical data structures, or application of this project's algorithms to more specific levels, such as the species or strain level, may provide a source for greater accuracy when classifying samples. Nevertheless, given the sensitivity of the gut microbiome to changes in diet (Wu et al., 2011) and some studies suggesting that the composition of the gut microbiome converges around three different enterotypes, or clusters, independent of geography, age, and genetics (Arumugam et al., 2011),

Table 4. Bayesian logistic regression weights associated with select microbial genera in the US/Venezuela (US/V) and Malawi/Venezuela (M/V) 959-feature binary classifiers. Large positive parameters correspond to features positively correlated with samples drawn from Venezuela, while negative weight parameters correspond to features positively correlated with samples drawn from the United States (in the first column) and Malawi (in the second column). Association with a Western diet (W) or vegetatrian diet (V) is based on (De Filippo, 2010); association with rural areas (R) is based on (Tyakht et al., 2013).

Genus	US/V	M/V	Association
Treponema	11.91	5.00	V
Lactobacillus	6.64	-0.32	
Prevotella	6.13	-1.64	V,R
Megasphaera	2.76	0.58	
Coptotermes	1.56	21.72	
Escherichia	-0.04	0.06	
Faecalibacterium	-0.05	1.46	R
Bifidobacterium	-0.66	-0.67	
Butyrivibrio	-0.75	18.35	V
Eubacterium	-1.12	5.77	
Hespellia	-1.34	11.72	
Veillonella	-2.19	-7.82	
Bacteroides	-2.77	1.26	W
Ruminococcus	-2.85	4.56	R
Akkermansia	-4.01	-3.55	
Clostridium	-4.33	-6.14	
Blautia	-4.86	3.65	
Parabacteroides	-8.69	28.75	W
Roseburia	-10.30	-32.80	R
Alistipes	-11.19	-9.03	
Dialister	-12.57	4.89	

accuracies close to 100% may not be achievable.

Finally, although the variability of the human microbiome over time may represent a source of uncertainty in some models, it is precisely because of this variability that human gut microbiome classifiers may be useful in characterizing the recent locations and interactions of people. Some studies have explored the role of abrupt changes in diet on the microbiome of individuals (Wu et al., 2011), and similar studies exploring the role of geographic relocation, illness, or cohabitation may provide other useful characterizations of activity.

In this project, we demonstrated the use of SVM and MAP-based classifiers to differentiate the geographic origin of human gut microbiome samples. Given the approximately 88-90% accuracies provided by classifiers in this project, an expanded set of such models may be useful for providing presumptive indicators of geographic origin in scientific and forensic studies.

References

[1] Arumugam, M, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346): 174–180, 2011.

[2] Backhed, F., et al., Host-bacterial mutualism in the human intestine. *Science*, 307:1915-1920, 2005.

[3] De Filippo, Carlotta, et al., Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *PNAS*, 107(33): 14691-14696, 2010

[4] Gelman, Andrew, et al., Data Analysis Using Regression and Multilevel/Hierarchical Models. The Comprehensive R Archive Network. Online, http://cran.r-project.org, 2013.

[5] Gunn, A, and Pitt, SJ, Microbes as forensic indicators. *Tropical Biomedicine*, 29(3): 311–330, 2012.

[6] Ling, Zongxin, et al., Pyrosequencing analysis of the human microbiota of healthy Chinese undergraduates. *BMC Genomics*, 14(390), 2013.

[7] Meyer, F., et al., The metagenomics RAST server a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386, 2008.

[8] Meyer, David, et al., e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. The Comprehensive R Archive Network. Online, http://cran.r-project.org, 2012.

[9] Nam, Young-Do, et al., Comparative Analysis of Korean Human Gut Microbiota by Barcoded Pyrosequencing. *PLoS ONE*, 6(7): e22109, 2011.

[10] Parks, DH, et al., GenGIS 2: Geospatial Analysis of Traditional and Genetic Biodiversity, with New Gradient Algorithms and an Extensible Plugin Framework. *PLoS ONE*, 8(7): e69885, 2013.

[11] Qin, Junjie, et al., A human gut microbial gene catalog established by metagenomic sequencing. *Nature*, 464(7285): 59-65, 2010.

[12] Tyakht, AV, et al., Human gut microbiota community structures in urban and rural populations in Russia *Nature Communications*, 4:2469, 2013.

[13] Wu, Gary D., et al. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science*, 334(6052): 105-108, 2011.

[14] Yatsunenko, Tanya, et al. Human gut microbiome viewed across age and geography. *Nature*, 486(7402): 222-227, 2012.