FORENSIC IDENTIFICATION WITH ENVIRONMENTAL SAMPLES

Gregory Ditzler and Gail Rosen[*]

Drexel University
Dept. of Electrical & Computer Engineering
Philadelphia, PA, 19104, USA

Robi Polikar[†]

Rowan University
Dept. of Electrical & Computer Engineering
Glassboro, NJ, 08028, USA

ABSTRACT

The field of forensics aims to understand the physical biomarkers that make each person unique. Recently, it has been discovered that one of the traits that makes us unique from one another are the composition of the microbial communities found throughout our bodies. For example, identical twins who share the same set of DNA may have vastly different microbial communities in or on various body sites. It was recently discovered that microbial communities can be exploited for forensic identification by clustering samples from individual's skin and objects that they may have previously touched. Typically, this is done by using basic multi-dimensional scaling analysis using phylogenetic distances. In this work, we circumvent the use of phylogenetic distances by using the raw community abundances, and we present an application of kernels for metagenomic data analysis. In addition, we show that strategic selection of features can improve classification accuracy.

Index Terms— forensics; metagenomics; bioinformatics

1. INTRODUCTION

Metagenomics is the study uncultured microorganisms obtained directly from an environmental sample [1]. In ecology, scientists are not only concerned about what species are in an environmental sample but how different samples compare. Through next generation sequencing, we are now able to collect, process and annotate sequences obtained from a microbe that contains thousands of microbial species, which can provide a plethora of information about the site from where the sample was obtained. Several recent studies have been conducted that examine the microbial communities of infant gut [2] and several body sites of adults [3]. Studies like these provide insight into how our microbial communities change over time and the effect that obesity, disease or cancer will have on our microbes. Perhaps of even greater concern for this work, is that our microbial communities maybe capable of telling us apart. Recent studies of the human microbiome have demonstrated that the microbial communities vary between individuals and between body sites [4]. Furthermore, recent studies into forensic applications of metagenomics have demonstrated that collecting environmental samples from individuals and objects may provide a significant insight into the identification of potential suspects [5]. Fierer et. al. collect metagenomic samples from three individuals' finger tips and the keyboards they have used, sequence the samples then apply UniFrac with a multi-dimensional scaling technique [5,6]. UniFrac is a phylogenetic distance computed using a phylogenetic tree generated from the metagenomic samples. The results indicate that the individual who last touched the keyboard can be identified, even hours after use.

The primary contributions of this work can be summarized as follows: (a) kernels are introduced for coordinate analysis schemes in ecology studies, (b) comparison of five different coordinate analysis methods for metagenomic feature extraction along with other popular feature extraction methods, (c) analysis of the clusterability of metagenomic samples for forensic identification, and (d) a preliminary analysis on the study of feature selection versus extraction for forensic identification.

2. METHODS FOR FORENSIC IDENTIFICATION

Related work in the study of the human microbiome have used standard coordinate analysis schemes with little justification of the methods of analysis. Therefore, in this work we use several coordinate analysis schemes along with a new implementation that takes advantage of kernels. It is important to note that while the microbial communities may be capable of telling us apart, there has been little work in the way of establishing a pipeline for forensic identification using environmental samples (e.g., preprocessing, feature extraction/selection, clusterability, classifier accuracy...). The process of addressing metagenomic samples is addressed in Fig. 1 and described in Section 2.1.

[*]G. Ditzler and G. Rosen are supported by the National Science Foundation (NSF) CAREER award #0845827, NSF Award #1120622, and the Department of Energy Award #SC004335. Author email: gregory.ditzler@gmail.com, gailr@ece.drexel.edu

[†]R. Polikar is supported by the NSF under Grant No: ECCS-0926159. Author email: polikar@rowan.edu

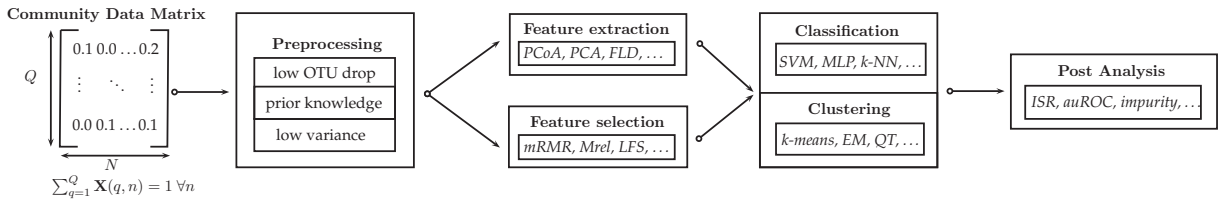


Fig. 1. Suggested pipeline for the use of metagenomic data derived from a community data matrix.

2.1. Analysis of Metagenomic Samples for Forensics

Our aim in this work is to investigate how to most accurately identify an individual solely based on a metagenomic sample by working with a community data matrix (CDM). The CDM is an $Q \times N$ matrix, where Q is the number of species, or operational taxonomic units (OTUs) and N is the number of samples. (We will use OTU instead of species in the rest of the paper, since OTU is a strict definition of species whereby the sampled 16S rRNA genes must be at least 97% similar to be considered within the same species.) Many metagenomic samples contain thousands of OTUs leading to very high dimensionality data sets; however, upon further investigation, we find that many of OTUs have a few occurrences. One approach to reduce the dimensionality, and hence to reduce the impact of the curse of dimensionality, is to remove OTUs that occur fewer times then a specified threshold. While the selection of such a threshold may seem rather arbitrary, it may be necessary to remove the OTUs that do not offer information, and may simply be viewed as a type of noise in the feature set.

An important question that we seek to answer from this analysis is whether features used to identify and group individuals should be extracted by means of a common transform or selected via a feature selection algorithm. On one hand, the feature selection methods provide biological meaning to the data (i.e., selection of attributes from the data), whereas feature extraction methods (i.e., derived via some transform or projection) may make attribute interpretation a bit ambiguous. For this work, we have selected several popular feature selection methods such as maximum relevance (Mrel) [7], maximum relevance minimum redundancy (mRMR) [7], linear forward search (LFS) [8], and a genetic search (GAFS) [9]. LFS and GAFS are implemented in the Weka data mining package [10]. For feature extraction, we test the Fisher linear discriminate (FLD), and principal coordinate analysis (PCoA) implemented with the Euclidean, Chord and Hellinger distances [11]. PCoA has been widely used in many ecological and metagenomic studies for multi-dimensional scaling. We also provide a variation of PCoA that uses kernels to provide the distance measures in feature space (refer to Section 2.2).

In order to address the question of feature selection versus extraction and which, if any preprocessing schemes are beneficial, we need to select a figure of merit to assess performance, or the identification success rate of the system. Thus, we need a way to measure closeness of the bacterial communities, so that we can explore the effect of using more/less principal coordinates. One figure of merit for the clusterability of the data would be to measure the impurity of the cluster (k-means is selected as the clustering algorithm). Our cluster impurity is given by:

$$\rho = 1 - \frac{1}{k} \sum_{\ell=1}^{k} \frac{\mid \mathcal{S}_{\omega_\ell} \cap \mathcal{C}_\ell \mid}{\mid \mathcal{C}_\ell \mid} \tag{1}$$

where $\mathcal{S}_{\omega_\ell} = \{(\mathbf{x}_n, y_n) \mid \mathbf{x}_n \in \mathcal{C}_\ell \text{ and } y_n = \omega_\ell\}$ is the set of data from class ω_ℓ assigned to cluster ℓ, $(\mathbf{x}_n, y_n)$ are data pairs, $\mathcal{C}_\ell$ is the set of data assigned to cluster ℓ, and ω_ℓ is the class that occurred most frequently in cluster ℓ. Essentially cluster impurity is the average error of each cluster, where the error is formed by associating each cluster with its most likely label. Data collected from the skin are used to initialize the clusters and data collected from the physical objects are then used for testing. Using the impurity measure allows us to report the clustering accuracy across varying numbers of coordinates retained after PCoA. Since the preprocessing and feature selection/extraction methods are to be applied to forensic identification, we selected a generic classifier for comparison – the support vector machine (SVM), which is not be limited by the curse of dimensionality. The SVM is used as the baseline classifier for this study due to its wide popularity, ability to solve non-linear problems via the kernel, and its computational efficiency with large scale implementation schemes [12]. The Shogun machine learning toolbox was used to implement the SVM [13]. Unless otherwise noted, all supervised methods are performed using only samples obtained from the skin and applied to samples obtained from objects (e.g., mice or keyboards).

2.2. Kernel Methods in Principal Coordinate Analysis

Kernels methods have shown great success in many areas of machine learning including classification, regression and component analysis [12, 14]. It only seems natural to apply kernel methods to PCoA techniques, as PCoA is used by many biologists and ecologists. In this section we provide the tools needed to integrate kernels with PCoA for data analysis.

Traditional PCoA methods perform multi-dimensional scaling on distance matrix, which measures the pairwise distance between samples [11]. Similar to principal component analysis (PCA), PCoA computes the eigenvectors and eigenvalues of the pairwise distance matrix. The principal

coordinates of the data are then derived from the eigenvectors. There are a large number of distance measures that can be applied to obtain the distance matrix, including phylogenetic distances, which require a phylogenetic tree to be available [6]. In this work, we use kernels as distance measures, which are computed as norms in feature space [14]. Traditional positive definite (pd) kernels provide us with a measure of similarity by way of the canonical dot product; however, a large class of kernels, known as conditionally positive definite (cpd), also exist for measuring dissimilarity rather than similarity in feature space. For example, a norm in feature space is calculated as:

$$\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\|^2 = k(\mathbf{x},\mathbf{x}) + k(\mathbf{x}',\mathbf{x}') - 2k(\mathbf{x},\mathbf{x}') \quad (2)$$

where $k(\mathbf{x},\mathbf{x}') = \Phi(\mathbf{x})^{\mathrm{T}}\Phi(\mathbf{x}')$ is the kernel. While there is a large body of work that describes the properties of cpd kernels [14], we simply note that the pairwise distance matrix used in PCoA can be computed using kernels that measure a norm distance in feature space. For this work we have selected the exponential and multi-quadratic kernels for the experiments as described in [14].

3. PRELIMINARY RESULTS

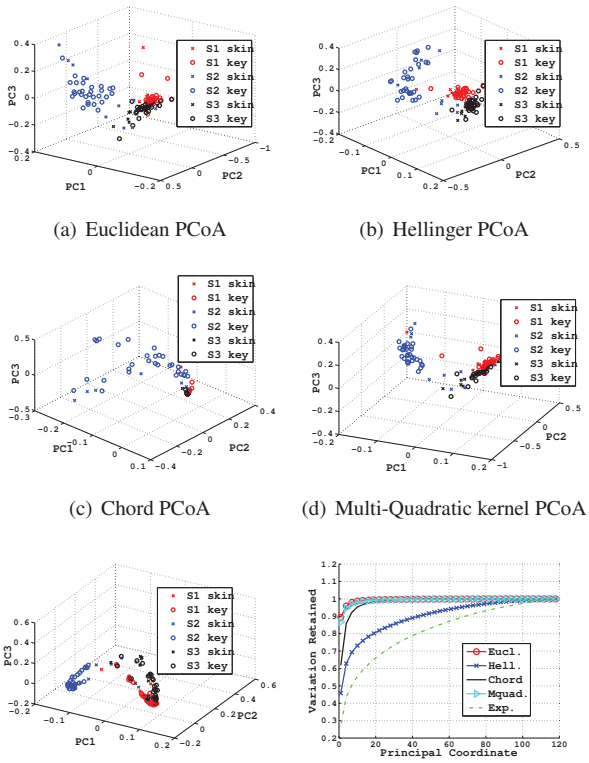
3.1. The Forensic Dataset

In this work we selected a real-world dataset, collected from the finger tips and keyboards of three individuals [5], which has been used as a benchmark in [15]. The environmental samples obtained from the keyboard were collected after several hours from the time the keyboard was initially touched by the subject. The original data consists of 120 samples from three subjects with approximately 3600 OTUs; however, we removed any OTU that has been detected less than 10 times for all 120 samples. Each sample is normalized prior to PCoA. We refer to this matrix as the community data matrix. For a more detailed description of the sample collection and sequencing of the forensic data, refer to [5].

3.2. Experiments

We begin discussing experimental results by applying PCoA schemes to the CDM. The three individuals are clearly differentiable using the metagenomic samples as shown in Fig. 2, which illustrates PCoA used with five different distance metrics. It appears, from visual inspection, that PCoA with the Hellinger distance provides the best separation of the individuals for the different distance metrics tested (refer to Fig. 2(b)). We also observe that the individual's skin and keyboard samples cluster together quite well, which was also observed in [5]. Finally, Fig. 2(f) contains the amount of variation in each principal coordinate for each distance metric tested. Note that while PCoA with the Euclidean distance provides the most variation in fewer coordinates, it is the Hellinger distance that provides the best differentiation between groups.

The data obtained from PCoA are clustered using k-means with the number of principal coordinates varied from



(a) Euclidean PCoA (b) Hellinger PCoA

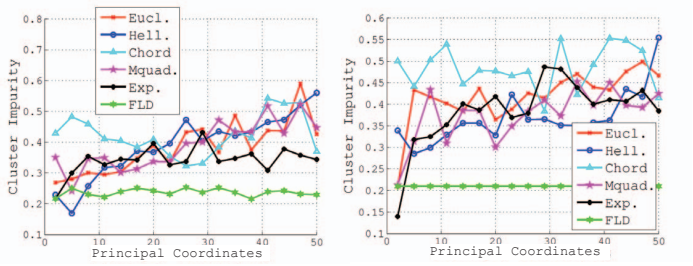
(c) Chord PCoA (d) Multi-Quadratic kernel PCoA

(e) Exponential kernel PCoA (f) Variation retained in the principal coordinates

Fig. 2. PCoA applied to the forensic data with three principal coordinates retained.

2 to 50. Clustering subjects is split into slightly different scenarios and the average cluster impurity is measured. As stated previously, this experiment is designed to demonstrate what is gained, or lost, due to dropping principal coordinates. The first scenario is the ability to cluster a subject with his/her own samples, regardless of the sample site (e.g., skin or keyboard). Two thirds of the data are used to initialize clusters and the remaining data are used for testing. Fifty independent trials are averaged to obtain the impurities in Fig. 3(a). The second scenario initializes the clusters with the skin data, and the clusters are evaluated on the data obtained from the keyboards. The cluster impurities for the second task are shown in Fig. 3(b).

A common trend observed in Fig. 3 is that increasing the number of coordinates does not necessarily reduce the cluster impurity. In fact, the impurity is increasing with the number of coordinates retained. It is clear that after five coordinates, there is not much to be gained in terms of improving the clusters, and that FLD is highly competitive with the PCoA methods performing at their best. We note that FLD's impurity contains very little variation, which can be attributed to random chance, as FLD reduces to two dimensions because the number of classes remains fixed at three (i.e., each person in



(a) Sample site independent (b) Sample site dependent

Fig. 3. Cluster impurities of k-means. (a) samples are randomly selected regardless of site to initialize and test k-means, and (b) samples from the skin are used to find the cluster centers and "label" each cluster then the keyboard data are used for testing.

Table 1. SVM ISR with a polynomial kernel. The multi-class SVM is implemented in a 1 vs. 1 as well as a 1 vs. all configuration.

Feature	1 vs. 1	1 vs. all	Feature	1 vs. 1	1 vs. all
Chord (2)	65.56	67.78	Exp. (2)	76.67	76.67
Chord (3)	51.11	51.11	Exp. (3)	76.67	71.11
Chord (5)	73.33	62.22	Exp. (5)	78.89	80.00
Chord (10)	75.56	66.67	Exp. (10)	68.89	73.33
Eucl. (2)	84.44	83.33	M-quad. (2)	76.67	74.44
Eucl. (3)	80.00	77.78	M-quad. (3)	82.22	74.44
Eucl. (5)	80.00	80.00	M-quad. (5)	75.56	74.44
Eucl. (10)	66.67	64.44	M-quad. (10)	73.33	70.00
Hell. (2)	84.44	81.11	FLD	**92.22**	91.11
Hell. (3)	91.11	85.56	LFS (9)	84.44	83.33
Hell. (5)	88.89	90.00	GAFS (1103)	83.33	84.44
Hell. (10)	80.00	78.89	Mrel (10)	88.89	85.56
CDM (444)	91.11	91.11	mRMR (10)	87.78	**96.67**
CDM (3695)	90.00	93.33			

the dataset).

Next, we address feature selection versus extraction. A multi-class SVM configuration was used with a polynomial kernel of order 6 and regularization of 100.0, so chosen to minimize the risk of over-fitting on such small data set. The SVM is trained using only the skin samples and tested on samples collected from a keyboard. Table 1 contains the identification success rate (ISR) of the SVMs for all features tested, where Chord, Eucl., Hell., Exp., & M-quad. are PCoA methods, CDM (444) are the original abundances with more than 10 occurrences, CDM (3695) are the original data with all OTUs, and mRMR, GAFS, & LFS are feature selection methods. The parentheses include the number of features. Using the raw CDM for this data set appears to be sufficient for good classification and removing samples due to low abundance. While blindly reducing the low OTU counts dramatically reduces the dimensionality, the reduced feature set has little impact on the ISR. The FLD appears to be a reliable feature extractor, but does not seem any more robust than using the raw feature set. However, the dimensionality with FLD is lower then that of the raw features. mRMR achieves the highest ISR, and appeared to be robust to kernel selection such as different order polynomials and RBF kernels (omitted due to space limitations). Of course, we must mention that these are very preliminary results and a more comprehensive analysis is required for the optimal selection of classifier, kernel, feature selection, and extraction.

4. CONCLUSIONS

In this work, we have started developing a general framework for reliable forensic identification using metagenomic samples. The methods test preprocessing and feature extraction/selection steps and the approach does not require a phylogenetic tree, which is required by the Unifrac distance [5,6]. We also show that while no previous metric can be declared the indisputable winner, using methods like FLD or mRMR consistently yields near-best results. The application of kernels for metagenomic data analysis was presented for PCoA. We have shown that feature extraction and selection via FLD, PCoA, or mRMR can provide improvement by at least 10% – a significant improvement – in the ISR of the system using a multi-class SVM. Future work will include the study of covariate shift and domain adaptation between the skin and keyboard samples, which poses additional hurdles for metagenomic forensics.

5. REFERENCES

[1] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics," *PLoS Comp. Bio.*, vol. 6, no. 2, pp. 1–13, 2010.

[2] J. E. Koenig et. al., "Succession of microbial consortia in the developing infant gut microbiome," *Proc. of the Nat'l Acad. of Scie.*, pp. 4578–4585, 2010.

[3] E. Costello et. al., "Bacterial community variation in human body habitats across space and time," *Science*, vol. 326, pp. 1694–1697, 2009.

[4] J. G. Caporaso et. al., "Moving pictures of the human microbiome," *Genome Biology*, vol. 12, no. 5, 2011.

[5] N. Fierer, C. Lauber, N. Zhou, D. McDonald, E. Costello, and R. Knight, "Forensic identication using skin bacterial communities," *Proc. of the Nat'l Acad. of Scie.*, vol. 107, no. 14, pp. 6477–6481, 2010.

[6] C. Lozupone and R. Knight, "UniFrac: a new phylogenetic method for comparing microbial communities," *Applied Environmental Microbiology*, vol. 71, no. 12, 2005.

[7] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information," *IEEE Trans. on Patt. Analy. and Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[8] M. Gütlein, E. Frank, and M. H. A. Karwath, "Large-scale attribute selection using wrappers," in *IEEE Symposium on Computational Intelligence and Data Mining*, pp. 332–339, 2009.

[9] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, 1989.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[11] J. Grower, "Multivariate analysis and multidimensional geometry," *Journ. of Royal Stat. Soc.*, vol. 17, no. 1, pp. 13–28, 1967.

[12] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verilag, 2nd ed., 1999.

[13] S. Sonnenburg et. al., "The SHOGUN machine learning toolbox," *Journal of Machine Learning Research*, vol. 11, pp. 1799–1802, 2010.

[14] B. Schlköpf and A. J. Smola, *Learning with Kernels*. The MIT Press, 1st ed., 2001.

[15] D. Knights, E. Costello, and R. Knight, "Supervised classification of human microbiota," *FEMS Microbiology Review*, vol. 35, pp. 343–359, 2010.