# Host Trait Prediction of Metagenomic Data for Topology-Based Visualization

Laxmi Parida[1,⋆], Niina Haiminen[1], David Haws[1], and Jan Suchodolski[2]

[1] Computational Biology Center, IBM T.J. Watson Research, Yorktown Heights, NY, USA
[2] Veterinary Medicine & Biomedical Sciences, Texas A&M University, College Station, TX, USA
parida@us.ibm.com

**Abstract.** Microbiome and metagenomic research continues to grow as well as the size and complexity of the collected data. Additionally, it is understood that the microbiome can have a complex relationship with the environment or host it inhabits, such as in gastrointestinal disease. The goal of this study is to accurately predict a host's trait using only metagenomic data, by training a statistical model on available metagenome sequencing data. We compare a traditional Support Vector Regression approach to a new non-parametric method developed here, called PKEM, which uses dimensionality reduction combined with Kernel Density Estimation. The results are visualized using methods from Topological Data Analysis. Such representations assist in understanding how the data organizes and can lead to new insights. We apply this visualization-of-prediction technique to cat, dog and human microbiome obtained from fecal samples. In the first two the host trait is irritable bowel syndrome while in the last the host trait is Kwashiorkor, a form of severe malnutrition.

## 1 Introduction

In recent years there has been an explosion of interest in microbiomes and metagenomics, which has been coupled with a dramatic increase in data to process and analyze. The microbiome is understood to be the community of microorganisms that inhabit some environment, such as the human gut, the soil surrounding plant roots, sewage treatment, etc. Metagenomics is the study of the genetic material of the microbes inhabiting some microbiome. Some studies focus on whole genomic sequencing of all organisms in the microbiome, providing massive amounts of data to analyze. Often though, many studies focus primarily on the diversity and specific abundance of each type of microorganism in one or many samples. To this end, sequencing typically targets the 16S rRNA gene which is present in most organisms. Equipped with 16S rRNA sequences, researchers are able to estimate which microorganisms are present in the environmental sample and classify the them from coarser to finer categories by

⋆ Corresponding author.

R. Natarajan et al. (Eds.): ICDCIT 2015, LNCS 8956, pp. 134–149, 2015.

phylum, class, order, family, and genus, with a loss of accuracy as one moves from coarser to finer classification.

It is understood that the microbiome plays a crucial role in the environment it inhabits, and may have a complicated relationship with the host or organism of interest. For example, the microbiome surrounding plant soil can have a dramatic effect on drought resistance [1], and conversely plants can effect the microbiome of the soil they inhabit [2] and changes in the health of a human host can directly impact the microbiome in the gut [3,4]. However, many microbiome collection efforts are focused on collecting samples from some environment and do not consider information about the environment or host, such as host disease status or other host phenotypes. As such, many studies and data sets contain an abundance of microbiome samples from multiple hosts, with little or no data on the host itself. Additionally, it is often difficult to understand and compare multiple microbiomes with respect to host traits. Nevertheless, we are interested in the relationship of a microbiome with respect to its associated host's traits.

The goal of this study is two-fold. First, to quantify the host status (e.g. disease status or some phenotype) by training a statistical model on available host and metagenomic data. From this, one can then attempt to predict a host's trait using only metagenomic data. This study is focused on Operational Taxonomic Unit (OTU) information for each microbiome and binary host traits. Second, provide a low-dimensional visualization of multiple host's microbiomes using tools from topological data analysis. The visualization is able to break down multiple microbiomes with respect to the host traits as well as highlight differences seen only at the microbiome level. The visualization is able to retain important structures of the high-dimensional data with the goal of leading to new insights and understanding of the otherwise opaque complicated data.

## 2 Methods

*Quantifying Host Trait.* All datasets that we studied contain multiple microbiome samples across multiple hosts. For each microbiome, OTU tables and the host's binary trait ($\{0, 1\}$) were obtained. The first step involves the training of a statistical model on available OTU and host trait data. Two approaches were used for this step, widely used Support Vector Regression (SVR) – a parametric approach – and a new non-parametric algorithm called **P**rediction through **K**ernel density **E**stimation of **M**etagenomic data, or PKEM for short. Both approaches are described below.

Support Vector Machines (SVMs) have been utilized in microbiome data analysis [5,6] due to their observed empirical performance on this type of data, as well as due to several theoretical considerations as summarized in [5]: SVMs perform well in data with limited sample size, are relatively insensitive to high dimensionality of the data, prevent overfitting by using regularization techniques, and can learn both simple and complex decision functions. Hence we also included SVMs as a state of the art method in our comparison.

The topological data analysis visualization then uses the above predictions of the host trait, as well as a distance measure between OTU tables. If available,

the weighted UniFrac [7] distance was used. The output is a low-dimensional representation of the microbiome data and is described below.

UniFrac distance is also used by existing microbial community analysis systems such as QIIME [8] that employs Random Forests for trait classification. Here we focus on SVM as the benchmark method for classification, for the above mentioned reasons.

*Support Vector Regression.* SVR [9,10,11,12,13] attempts to model the relationship between the explanatory and response variables by finding a hyperplane (high-dimensional generalization of a 3d-plane), where all the data points lay either on the hyperplane or as close as possible to it. The real trick here is that the data are first mapped to a different high-dimensional space using possibly a non-linear kernel.

Following [14], given a training set $(\mathbf{x}_i, y_i)$, $i = 1, \ldots l$, where $\mathbf{x}_i \in \mathbb{R}^n$, the goal of $\varepsilon$-SV regression is to find a function $f(\mathbf{x})$ that is at most $\varepsilon$ deviation from the explanatory variable $y_i$ over the response variable $\mathbf{x}_i$, while remaining as flat as possible in the feature space. In our case, the response variables will be OTU data, and the explanatory variable will be the host trait associated with the microbiome. Training an SVR requires solving

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\left(\sum \xi_i + \sum \xi_i^*\right) \\ &\text{subject to} \begin{cases} y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \le \varepsilon + \xi_i \\ \mathbf{w}^\top\phi(\mathbf{x}_i) + b - y_i \le \varepsilon + \xi_i^* \end{cases} \\ &\xi_i, \xi_i^* \ge 0. \end{aligned} \tag{1}$$

The data vectors $\mathbf{x}_i$ are mapped to another space via the function $\phi$, and SVR attempts to fit the data in this higher dimensional space. Thus, the choice of $\phi$, referred to as the *kernel*, has a large impact. The de-facto SVR software `libsvm` [15] provides four kernels:

$$\begin{aligned} \text{Linear:} \quad & \mathbf{u}^\top\mathbf{v}, \\ \text{Polynomial:} \quad & (\gamma\mathbf{u}^\top\mathbf{v} + r)^d,\ \gamma > 0, \\ \text{Radial:} \quad & exp(-\gamma\|\mathbf{u} - \mathbf{v}\|^2),\ \gamma > 0, \\ \text{Sigmoid:} \quad & tanh(\gamma\mathbf{u}^\top\mathbf{v} + r). \end{aligned}$$

Conversely, *Support Vector Machine* (SVM) attempts to find a hyperplane separating a set of data points and is used for binary classification. In this case the inequalities in Equation 1 are reversed and thus data points are penalized for being too close to the separating hyperplane via the $\xi$ and $\xi^*$ parameters appearing the cost function. The CRAN `e1071` [16] R [17] package was used for all SVR and SVM computations.

*PKEM.* A second non-parametric prediction method was developed called **P**rediction through **K**ernel density **E**stimation of **M**etagenomic data, or PKEM

for short. It combines a dimensionality reduction step with multivariate kernel density estimation. The dimensionality reduction step is often required since kernel density estimation can lead to improper fitting of the data when sample sizes are small relative to the dimension of the data. Classical principal component analysis is used for the dimensionality reduction step.

Principal Component Analysis (PCA) [18,19] is a well-established method which uses orthogonal transformations such that the first principal component contains the largest variance, the second principal component contains the second largest variance, and so on. PCA is often used on high-dimensional data to transform and truncate the data to a lower dimensional space, while attempting to preserve as much variance as contained in the original data. That is, PCA reformulates the data according to the principal components, ranking from most important to least important. By truncating the least important principal components, one retains the most important parts of the original data while reducing its dimension.

PCA can be accomplished using a singular value decomposition. Any real $m \times n$ matrix $M$ can be written $M = U\Sigma V^\top$. Here, $U$ is an $m \times m$ *unitary* ($U^\top U = UU^\top = I$) matrix, $\Sigma$ is an $m \times n$ rectangular diagonal matrix containing the singular values from largest (upper left) to smallest (lower right), and $V^\top$ is the transpose of an $n \times n$ unitary matrix. The PCA transformation is given by $U\Sigma$, and the $p$th PCA truncation is given by $U_p\Sigma_p$ where $U_p$ and $\Sigma_p$ are the first $p$ rows of $U$ and $\Sigma$ respectively.

Kernel density estimation (KDE) [20,21] is a non-parametric approach to estimate the probability distribution of a random variable. That is, if one has a sample of a random variable, kernel density estimation can be used to find an approximation of the unknown distribution underlying the random variable. Conceptually, kernel density estimation is similar to a histogram of the sample data, but with a smoothing out operation.

In the univariate case, if $(x_1, \ldots, x_n)$ are sampled identically and independently from a distribution with some unknown density function $f$, the goal of kernel density estimation is to estimate $f$ via some function $\hat{f}_h$. It does this by giving a little bit of weight to each sample and is formulated as

$$\hat{f}_h(x) := \frac{1}{nh} \sum_{i=1}^{n} K(x - x_i).$$

The function $K(\cdot)$ is the *kernel* and it is assumed to be symmetric and integrates to 1. The parameter $h$ is called the *bandwidth* and is chosen as small as the data will allow. Typically a Gaussian kernel is used for $K(\cdot)$. One way to visualize kernel density estimation is to imagine that for each data point on the real line, a handful of dirt is dropped (which makes a nice Gaussian dirt hill). Thus, if a group of data points are close on the real line, then a large mound of dirt accumulates around the group of data points since one dropped many handfuls of dirt around there. See Figure 1 for an example.

Multivariate kernel density [22] estimation is nearly identical to the above univariate case, except the kernel is almost always a multivariate Gaussian and

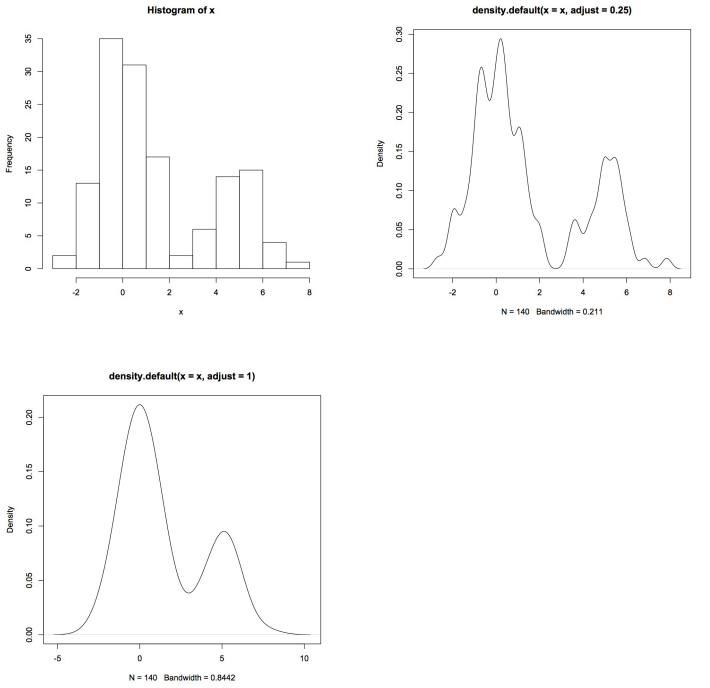


**Fig. 1.** Histogram of 140 data points (top left) sampled from some unknown distribution. Kernel density estimation using 0.25 of normal bandwidth (top right). Kernel density estimation using normal bandwidth (bottom).

the bandwidth parameter $h$ is replaced by a bandwidth matrix $H$ which is symmetric and positive definite. The bandwidth matrix $H$ determines the shape of the multivariate Gaussian kernel $K(\cdot)$.

The PKEM algorithm can be summarized:

1. Let $\mathbf{X}$ be a matrix where the rows are the OTU fractions and the columns are the $N$ hosts being studied.
2. Let $\mathbf{Y}$ be the $\{0, 1\}$ host traits.
3. Let $p$ be the user-input truncation dimension.
4. Perform PCA on the subsets of columns of $X$ with host trait $Y$ equal to 0 (or 1). Use obtained PCA transformation on all data $X$, call transformed data $B$.
5. Train a multivariate kernel density estimation function $F$ using columns of $B$ with host trait $Y$ equal to 0 (or 1).
6. Output kernel density estimation function $F$ which takes as input any OTU table data and outputs an estimate of its density estimation of the associated host trait to be 0 (or 1).

The PKEM algorithm can be trained on either the OTU data taken from hosts with traits valued 0 or valued 1, which may lead to different results depending on

the data. The `prcomp` function in the `stats` base package of R [17] was used to compute PCA and the `np` [23] R package was used to perform multivariate kernel density estimation. Additionally, the function $F$ output by KDE is normalized by the largest estimated value, yielding values in $[0, 1]$.

### 2.1 Topological Data Analysis

Topology is the mathematical study of spaces and their qualities, such as properties of spaces that are preserved under continuous deformations. Topological Data Analysis (TDA) is the application of the mathematically rigorous field of algebraic topology towards understanding large and high-dimensional data. Recently there has been a rapid growth in interest in TDA and its many applications [24,25,26,27,28].

One exciting application of TDA is forming reliable low-dimensional representations of high-dimensional data, with the ambition that the low-dimensional representation maintains important relationships and can be easier to interpret, leading to new insights on the otherwise opaque high-dimensional data. One recently introduced popular approach called `mapper` [29] has been successfully applied in many additional studies [28,26]. The fundamental concept of `mapper` is that the output is a combinatorial graph, as opposed to a set of data points or some subspace. Additionally, TDA is more robust to noise, can handle large data sets, and can handle any notion of distance. That is, one does not need to use Euclidean distance of data and may choose a more appropriate measure of distance. Lastly, `mapper` requires a *filter function* on the high-dimensional data, which is some real-valued function. That is, the filter function $f$ assigns some real value for each high-dimensional data point.

The `mapper` approach works roughly as follows: The filter function is applied to the input high-dimensional data $\mathbf{X}$ and the filter values are saved. Then the range of filter values are divided up into $k$ overlapping intervals. For each interval of filter values, the subset of data from $\mathbf{X}$ corresponding to the current filter interval is clustered. This clustering is performed for each filter interval. Once all the clusters have been formed, a graph is drawn with a node for each cluster. Since the filter intervals were overlapping, two clusters may share a data point from $\mathbf{X}$ in common. Thus, if two nodes (clusters) share at least one data point then an edge is drawn between the nodes. This completes the original `mapper` algorithm, but additional visualization can be performed such as coloring each node based on the average filter values as well as plotting any additional meta data about the data points in each node. For an example of `mapper` see Figure 2.

It must be emphasized that the output of `mapper` is highly dependent on the filter function chosen, the amount of overlap in the filter intervals, and the distance used for clustering as well as the clustering method. However, the output low-dimensional representation will often reflect the properties of the original high-dimensional data.

For this study, the PKEM and SVR algorithms were used to compute filter values. If the host trait was disease (0 healthy, 1 disease), than each algorithm estimates a host's disease status from 0 to 1 depending on its associated microbiome

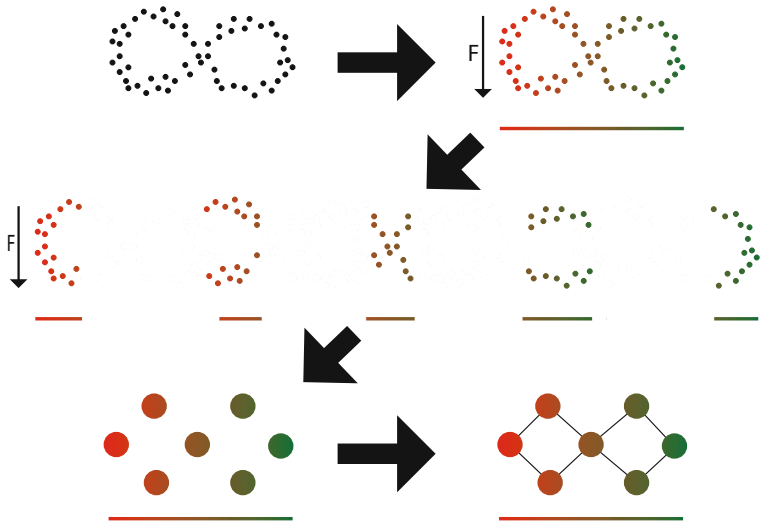


**Fig. 2.** Example of `mapper` algorithm. First a filter function $F$ is applied to the data. In this case, data are given a high value if they are to the left (red) and a low value if they are to the right (green). Second, the range of filter values are formed into overlapping intervals, creating corresponding collections of the original data. Third, each collection of data is independently clustered. Lastly, an edge is drawn between two clusters if they have at least one element of the data in common.

OTU data. The distance between OTU tables was either the weighted UniFrac distance if available, or the Euclidean distance. Clustering was performed using hierarchical clustering and the Ward method. Further, the number of clusters was determined in an unsupervised way by choosing the number of clusters which maximized the mean silhouette score [30].

### 2.2 Data

**Cat and Dog Data.** All samples were from dogs and cats that lived in home environments and were collected by veterinarians who evaluated the animals for their GI disease. Healthy animals were owned by students and staff at Texas A&M University. All samples were stored frozen at -80°C until processing of samples for DNA extraction. A 100mg (wet weight) aliquot of feces was extracted by a bead-beating method using a commercial DNA extraction kit (ZR Fecal DNA Kit$^{TM}$, Zymo Research Corporation) following the manufacturer's instructions. The bead beating step was performed on a homogenizer (FastPrep-24, MP Biomedicals) for 60 s at a speed of 4 m/s. The collection and analysis of fecal samples was approved by the institutional Clinical Research Review Committee of the college of Veterinary Medicine, Texas A&M University.

*Cat Data.* Fecal samples were obtained from healthy cats ($n = 23$) and cats with diarrhea ($n = 76$). Diseased cats were further compared based on the duration of their diarrhea: duration $< 21$ days ($n = 32$) vs. duration $> 21$ days ($n = 44$). None of the animals received antibiotics within 3 months of sample collection. Sequencing was performed targeting the V4 region of the 16S rRNA gene using forward and reverse primers: 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 806R (5'- GGACTACVSGGGTATCTAAT-3') using the Ion Torrent platform at a depth of 15,000 sequencing reads per sample. Operational taxonomic units (OTUs) were assigned based on at least 97% sequence similarity using QIIME v1.7. The sequences were deposited in SRA under accession number SRP047088. A similar data collection and analysis process is described also elsewhere [31].

*Dog Data.* Fecal samples were collected from healthy dogs ($n = 98$), dogs with chronic enteropathy (IBD, $n = 79$), and dogs with acute hemorrhagic diarrhea ($n = 15$). All dogs with CE were evaluated by endoscopic examination and intestinal inflammation was confirmed by histopathology. Dogs with acute diarrhea were worked up for the GI disease and were all diagnosed with uncomplicated diarrhea that resolved with routine symptomatic treatment within one week of presentation. None of the animals received antibiotics within 3 months of sample collection. Sequencing was performed targeting the V4 region of the 16S rRNA gene using forward and reverse primers: 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 806R (5'- GGACTACVSGGGTATCTAAT-3') using the Illumina platform at a depth of 5,000 sequencing reads per sample. Operational taxonomic units (OTUs) were assigned based on at least 97% sequence similarity using QIIME v1.7. A similar data collection and analysis process is described also elsewhere [32].

**Kwashiorkor Data.** Publicly available data was taken from a study of gut microbiomes of Malawian twins suffering from Kwashiorkor, a form of sever acute malnutrition [33] [1]. In the study, 317 Malawian twin pairs were followed for three years during which 43% became discordant (Kwashiorkor). In such discordant cases, both twins were fed ready-to-use therapeutic food (RUTF). The authors of the above study observed that the consumption of RUTF by discordant individuals eventually led to an improved health of the individuals' microbiome, and if RUTF was stopped prematurely the microbiomes regressed to their discordant state. Additionally, when the authors transplanted discordant microbiomes into gnotobiotic mice and provided a Malawian diet, the kwashiorkor microbiome lead to drastic weight loss as well as changes in their metabolism.

Phylum, class, order, family, and genus level 16S OTU data was taken from the original study for all individuals. Additional data was included here, specifically *weight-for-height* z (WHZ) score, RUTF consumption, and age. Multiple microbiome samples were available for each individual, each labeled with the state *healthy* or *kwashiorkor*. All *kwashiorkor* samples were included in the anal-

[1] Data retrieved from Jeffrey Gordon website:
`http://gordonlab.wustl.edu/SuppData.html`

ysis presented here, however, if an individual had multiple *healthy* microbiome samples, only the sample with the highest WHZ score was included.

## 3 Results

### 3.1 Prediction Accuracy

The ability of SVR, SVM, and PKEM to accurately predict the host's trait was tested by tenfold cross validation. That is, OTU and host trait data were split into ten evenly sized sets. Then SVR, SVM, and PKEM were trained on 90% of the available data and each method was used to predict the remaining 10% of the data. Both SVR and PKEM can be coerced to output a $[0, 1]$ continuous estimate of the host's binary trait. Thus a threshold is used to determine if the predicted host trait is 0 or 1. Figure 3 show the false positives vs. true positives as the threshold for the $\{0, 1\}$ classification varies. The parametric SVR outperforms the non-parametric PKEM. However, PKEM does remain viable as a classifier, as long as the threshold is low (approximately 0.2–0.3). Note, the linear, polynomial, and sigmoid SVR kernels were also studied but did not perform as well as the radial kernel.

Thresholds were set to 0.25 for PKEM and 0.50 for SVR and the F-score and accuracy of SVR, SVM, and PKEM were computed, see Table 1. Accuracy is reported in term of the *F-score* (F1) and the *accuracy* (ACC). Let TP=True Positives, TN=True Negatives, FP=False Positive, FN=False Negatives, P=Positive instances, and N=Negative instances, then $F1 := 2TP/(2TP + FP + FN)$ and $ACC := (TP + TN)/(P + N)$. Notice that the F1 score is primarily influenced by the $TP$.

In the Dog data, PKEM (truncation dimension 6) is comparable to SVR (linear) and SVM (linear) in terms of best accuracy 0.60. However, PKEM suffers from fewer TP, and thus has a lower F1 score. SVM slightly outperforms SVR in terms of F1, but not by a large margin, likely due to a poor choice of threshold for SVR (0.50).

In the Cat data, SVR (radial) and SVM (radial) have highest accuracy (0.81), while PKEM (truncation dimension 6) is slightly behind (0.78). Again, PKEM under performs in the F1 score due to low TP, although it has better F1 score than SVR (linear, polynomial) and SVM (polynomial, radial). In this case, SVM seems to suffer from very low TP and thus low F1 scores.

In the Kwashiorkor data, SVM (radial, sigmoid) ties for highest accuracy (0.64) and SVM (radial) has the highest F1 (0.76). The Kwashiorkor data set presents the largest difference in F1 and accuracy between SVM versus SVR and PKEM. It is good to note that, in this case, SVR and PKEM perform similarly.

Across all data, SVM attains the highest accuracy, or is at least as good as SVR and PKEM. Although, SVM performs poorly in terms of F1 score on the cat data. It is clear that all methods are able to use OTU microbiome data alone in order to predict the host's trait value. Additionally, the linear and radial kernel often perform best. For this reason, the radial kernel was chosen for use in the TDA visualization. PKEM performs well in terms of accuracy, although it does

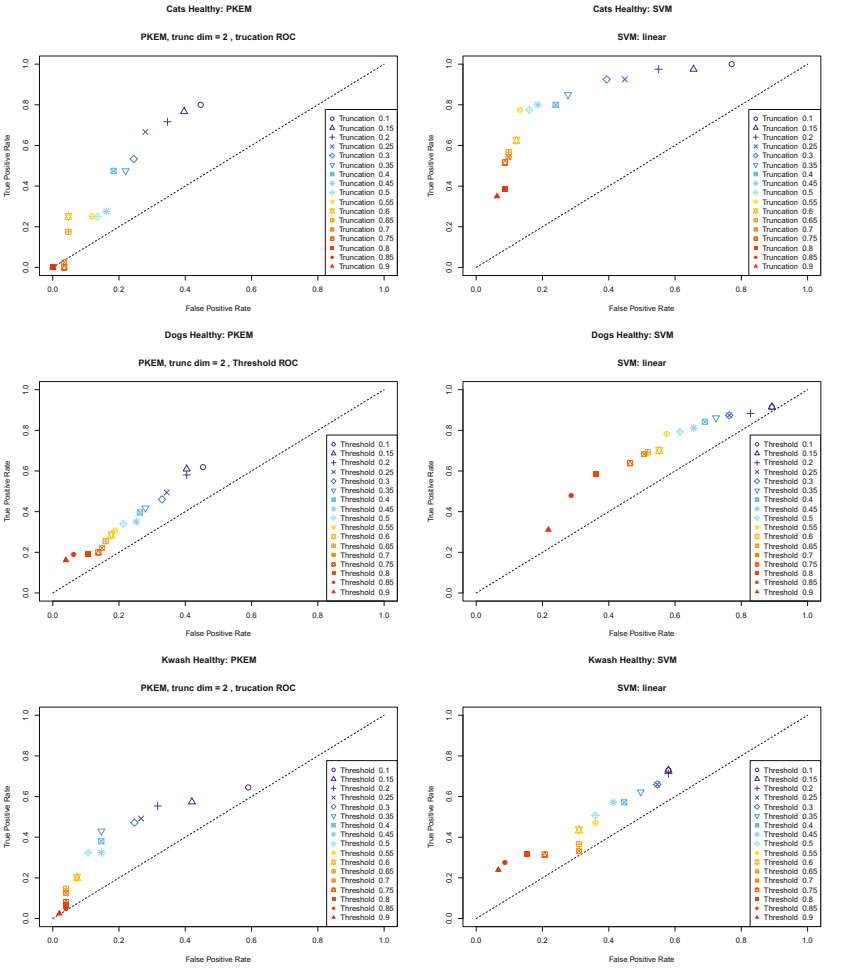


**Fig. 3.** Receiver of Operator Curves (ROC). The threshold to decide the $\{0, 1\}$ host trait was varied and the false positive vs. true positive rates were recorded. An ideal classifier would have ROC points in the upper left. The diagonal reflects a random classifier.

fall behind in terms of F1 score due to low TP. However, PKEM does not suffer from a choice of a kernel, since it is non-parametric.

Although SVM has high accuracy, the fact that it gives a binary $\{0, 1\}$ classification does not allow its use in the Topological Data Analysis described herein, whereas SVR and PKEM both give an estimate of the host trait value in the range of $[0, 1]$.

**Table 1.** Mean *F-score* (F1) and *accuracy* (ACC) of SVR, SVM, and PKEM on Dogs, Cats, and Kwashiokor data under ten-fold cross validation. Let TP=True Positives, TN=True Negatives, FP=False Positive, FN=False Negatives, P=Positive and N=Negative instances, then $F1 := 2TP/(2TP + FP + FN)$ and $ACC := (TP + TN)/(P + N)$. PKEM discrimination threshold set to 0.25. SVR threshold set to 0.50.

| | | Dog | | Cat | | Kwashiorkor | |
|---|---|---|---|---|---|---|---|
| | | F1 | ACC | F1 | ACC | F1 | ACC |
| SVR | Linear | 0.68 | **0.60** | 0.41 | 0.58 | 0.49 | 0.51 |
| | Polynomial | 0.68 | 0.56 | 0.38 | 0.68 | 0.58 | 0.53 |
| | Radial | 0.66 | 0.58 | **0.66** | **0.81** | 0.58 | 0.57 |
| | Sigmoid | 0.66 | 0.59 | 0.57 | 0.76 | 0.60 | 0.57 |
| SVM | Linear | **0.69** | **0.60** | 0.58 | 0.83 | 0.73 | 0.65 |
| | Polynomial | **0.69** | 0.54 | 0.07 | 0.76 | 0.74 | 0.62 |
| | Radial | **0.69** | 0.54 | 0.33 | **0.81** | **0.76** | **0.64** |
| | Sigmoid | **0.69** | 0.54 | 0.49 | 0.83 | 0.74 | **0.64** |
| PKEM | Trunc Dim 2 | 0.59 | 0.56 | 0.43 | 0.52 | 0.50 | 0.44 |
| | Trunc Dim 4 | 0.56 | 0.58 | 0.45 | 0.69 | 0.56 | 0.57 |
| | Trunc Dim 6 | 0.53 | **0.60** | 0.46 | 0.78 | 0.56 | 0.60 |
| | Trunc Dim 8 | 0.47 | 0.57 | 0.24 | 0.76 | 0.46 | 0.55 |
| | Trunc Dim 10 | 0.38 | 0.54 | 0.10 | 0.78 | 0.32 | 0.49 |

### 3.2 Topological Data Analysis on Metagenomic Data

Topological data analysis, specifically the `mapper` algorithm, was applied to the Cat, Dog, and Kwashiorkor data. Recall that `mapper` requires as input a set of filter values, the number of overlapping intervals to break the filter values into, the percentage overlap of each interval, a pairwise distance between points, and a clustering method. The SVR (radial) and PKEM algorithms were trained on each data set, where each algorithm was trained on the entire data set and the prediction was used as the filter values. The number of intervals was set to six with an overlap of 90%.

For clustering, the hierarchical method was used with Ward criteria for joining two clusters, which merges two clusters that minimize the resulting within-cluster variance. Here an unsupervised approach was taken by cutting the hierarchical clustering dendogram at 1.0, which in effect does hierarchical clustering using the Ward criteria and merges clusters as long as the within-cluster variance does not exceed 1.0. In the case of the Cat and Dog data the weighted UniFrac distance was used as input to the `mapper` algorithm in order to perform the cluster analysis. Whereas in the Kwashiorkor data the Euclidean distance between OTU samples was used as the distance measure.

Figure 4 shows the output of `mapper` applied to the Cat, Dog, and Kwashiorkor data. Additionally recall that `mapper` connects two nodes (two clusters) if they share an individual in common. In Figure 4 the number of overlapping individuals is given on the edge in orange. Lastly, summaries of metadata for each dataset

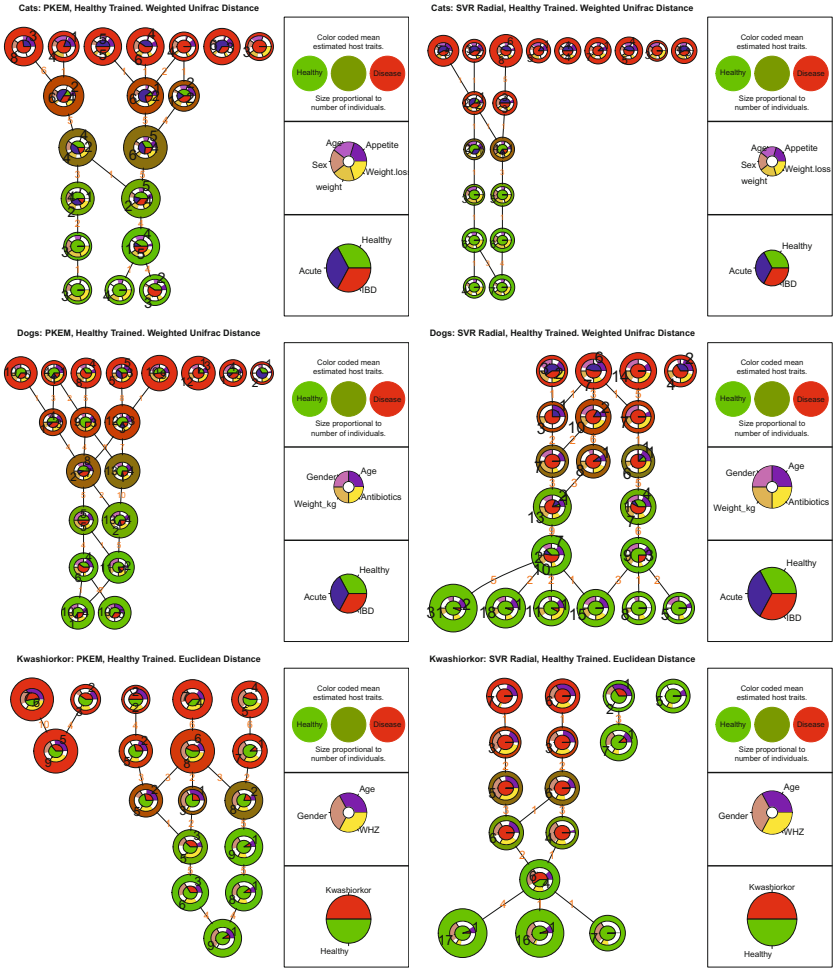


**Fig. 4.** Output of `mapper` applied to the Cat, Dog, and Kwashiorkor data. Left, the PKEM algorithm is used to compute the filter values where on the right the SVR algorithm with a radial kernel is used. Both used the weighted UniFrac distance.

and each cluster are also presented in Figure 4 and the details of each figure are discussed below.

The TDA of the Cat data using PKEM is given in the upper left of Figure 4. Inner curved bar plots give normalized mean Age, Appetite, Sex, Weight, and Weight Loss. In the bottom a bifurcation of the healthy-like cats appears where on the left the individuals appear to have higher weight and higher weight loss. The healthy-like individuals on the right contain some mis-classified individuals with IBD. The middle portion of clusters that are between healthy and disease also shows a splitting of the data, where sex and appetite may play a role.

The TDA of the Cat data using SVR (radial) is given in the upper right of Figure 4, with the same meta-data as above. In this case there are few misclassifications in terms of IBD and acute. However, the most healthy-like individuals do not seem to distinguish much in terms of the given meta data. However, for the middle portion of the graph the disease-like to healthy-like data seems to separate into two connected lines primarily by sex.

The TDA of the Dog data using PKEM is given in the second row and left of Figure 4. Inner curved bar plots give normalized mean Age, Antibiotics, Weight, and Gender. As in the Cat data using PKEM, there are some misclassification of IBD and acute. On the bottom a bifurcation can be seen of the healthy-like individuals noticeably by the application of Antibiotics or not and Age. In the second row from the top of the clusters, there are three disease-like clusters where it appears the cluster in the center distinguishes itself from the other two by Weight.

The TDA of the Dog data using SVR (radial) is given in the second row and right of Figure 4. In this case the healthy-like clusters are more abundant, but with some more misclassification compared to the Cat data using SVR (radial). For these healthy-like clusters, the distinguishing information seems to be the Age and Gender of the individuals involved. Additionally, the center healthy-like cluster has a high use of antibiotics. The second and third row of disease-like clusters show a partitioning of the data distinctly by Gender, and the use of Antibiotics or not.

The TDA of the Kwashiorkor data using PKEM is given in the last row and left of Figure 4. Inner curved bar plots give normalized mean Age, Gender, and WHZ. Again, there are some misclassification in the case of PKEM. Additionally in this case, the most healthy-like individuals cluster together. The third row down of clusters of in between healthy-like and disease-like appears to cluster first by Age (right cluster) and then the remaining two clusters appear to distinguish from one another by WHZ. In the case of the disease-like clusters they appear to distinguish from one another by either Age, WHZ, or Gender indicating each may be an important factor in the composition of the microbiome.

The TDA of the Kwashiorkor data using SVR (radial) is given in the last row and right of Figure 4. In this case, the healthy-like individuals are quite separated into multiple clusters where Gender and Age may play the biggest role. In the middle section of clusters in between healthy-like and disease-like there is a bifurcation of individuals primarily by Age and a Gender. In the case of the most disease-like surprisingly Gender, and to a lesser degree WHZ, appears to be a large factor in distinguishing microbiomes

## 4 Discussion

Research into microbiomes and metagenomics will only continue to grow, as well as the size and complexity of the available data. Additionally, the connection between host traits and the microbiome is only beginning to be elucidated and needs further study. We demonstrated here that, in fact, statistical models can

be trained on OTU metagenomic data and applied to accurately predict host traits.

Gastrointestinal disease is most likely a combination of various environmental factors and therefore it is not possible to define a clear cut host trait. Thus while feces would have lower sensitivity for separation, intestinal biopsies may have a much higher rate. This is also corroborated in Crohn's disease [34] where the authors observed a lower sensitivity when classifying disease status using fecal samples (See Figure 4) as compared to using tissue samples.

Finally there is a need to visualize and understand the ever-growing complex metagenomic data. Combining traditional prediction algorithms or novel non-parametric prediction methods such as PKEM with powerful topological data analysis can lead to improved insights into the data. For example, visualizing the data along with annotations such as antibiotic usage, age, and weight can assist in understanding the separation between healthy and afflicted individuals.

## References

1. Zolla, G., Badri, D.V., Bakker, M.G., Manter, D.K., Vivanco, J.M.: Soil microbiomes vary in their ability to confer drought tolerance to Arabidopsis. Applied Soil Ecology 68, 1–9 (2013)
2. Badri, D.V., Quintana, N., El Kassis, E.G., Kim, H.K., Choi, Y.H., Sugiyama, A., Verpoorte, R., Martinoia, E., Manter, D.K., Vivanco, J.M.: An ABC transporter mutation alters root exudation of phytochemicals that provoke an overhaul of natural soil microbiota. Plant Physiology 151(4), 2006–2017 (2009)
3. Devaraj, S., Hemarajata, P., Versalovic, J.: The human gut microbiome and body metabolism: implications for obesity and diabetes. Clinical Chemistry 59(4), 617–628 (2013)
4. Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C., Ley, R.E.: A guide to enterotypes across the human body: Meta-analysis of microbial community structures in human microbiome datasets. PLoS Computational Biology 9(1), e1002863 (2013)
5. Statnikov, A., Alekseyenko, A.V., Li, Z., Henaff, M., Perez-Perez, G.I., Blaser, M.J., Aliferis, C.F.: Microbiomic signatures of psoriasis: Feasibility and methodology comparison. Scientific Reports (3) (2013)
6. Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., Pei, Z., Blaser, M., Aliferis, C., Alekseyenko, A.: A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome 1(1) (2013)
7. Lozupone, C., Knight, R.: UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71(12), 8228–8235 (2005)
8. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al.: QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7(5), 335–336 (2010)
9. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144–152 (1992)

10. Guyon, I., Boser, B., Vapnik, V.: Automatic capacity tuning of very large VC-dimension classifiers. Advances in Neural Information Processing Systems, 147–155 (1993)
11. Cortes, C., Vapnik, V.: Support-vector networks. In: Machine Learning, pp. 273–297 (1995)
12. Schölkopf, B.: Support vector learning (1997), http://www.kernel-machines.org
13. Vapnik, V., Golowich, S.E., Smola, A.: Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems 9, 281–287 (1996)
14. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing 14(3), 199–222 (2004)
15. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
16. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., Weingessel, A.: e1071: Misc Functions of the Department of Statistics (e1071), TU Wien (2011) R package version 1.6
17. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2014)
18. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24(6), 417 (1933)
19. Pearson, K.: LIII. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2(11), 559–572 (1901)
20. Parzen, E.: On estimation of a probability density function and mode. The Annals of Mathematical Statistics, 1065–1076 (1962)
21. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics 27(3), 832–837 (1956)
22. Simonoff, J.S.: Smoothing methods in statistics. Springer, London (1996)
23. Hayfield, T., Racine, J.S.: Nonparametric econometrics: The np package. Journal of Statistical Software 27(5) (2008)
24. Zomorodian, A., Carlsson, G.: Computing persistent homology. Discrete & Computational Geometry 33(2), 249–274 (2005)
25. Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society 46(2), 255–308 (2009)
26. Nicolau, M., Levine, A.J., Carlsson, G.: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences 108(17), 7265–7270 (2011)
27. Chan, J.M., Carlsson, G., Rabadan, R.: Topology of viral evolution. Proceedings of the National Academy of Sciences 110(46), 18566–18571 (2013)
28. Bartlett, C.W., Cheong, S.Y., Hou, L., Paquette, J., Lum, P.Y., Jäger, G., Battke, F., Vehlow, C., Heinrich, J., Nieselt, K., et al.: An eQTL biological data visualization challenge and approaches from the visualization community. BMC Bioinformatics 13(suppl. 8), S8 (2012)
29. Singh, G., Mémoli, F., Carlsson, G.E.: Topological methods for the analysis of high dimensional data sets and 3D object recognition. In: SPBG, pp. 91–100 (2007)
30. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987)

31. Bell, E.T., Suchodolski, J.S., Isaiah, A., Fleeman, L.M., Cook, A.K., Steiner, J.M., Mansfield, C.S.: Faecal microbiota of cats with insulin-treated diabetes mellitus. PLoS ONE 9(10) (2014)
32. Suchodolski, J.S., Markel, M.E., Garcia-Mazcorro, J.F., Unterer, S., Heilmann, R.M., Dowd, S.E., Kachroo, P., Ivanov, I., Minamoto, Y., Dillman, E.M., Steiner, J.M., Cook, A.K., Toresson, L.: The fecal microbiome in dogs with acute diarrhea and idiopathic inflammatory bowel disease. PLoS ONE 7(12) (2012)
33. Smith, M.I., Yatsunenko, T., Manary, M.J., Trehan, I., Mkakosya, R., Cheng, J., Kau, A.L., Rich, S.S., Concannon, P., Mychaleckyj, J.C., Liu, J., Houpt, E., Li, J.V., Holmes, E., Nicholson, J., Knights, D., Ursell, L.K., Knight, R., Gordon, J.I.: Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. Science 339(6119), 548–554 (2013)
34. Gevers, D., Kugathasan, S., Denson, L.A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., Yassour, M., et al.: The treatment-naive microbiome in new-onset Crohns disease. Cell Host & Microbe 15(3), 382–392 (2014)