Feature Subset Selection for Inferring Relative Importance of Taxonomy

Gregory Ditzler*
Drexel University
Dept. of Electrical & Computer Engineering
Philadelphia, PA 19104 USA
gregory.ditzler@gmail.com

Gail Rosen
Drexel University
Dept. of Electrical & Computer Engineering
Philadelphia, PA 19104 USA
gailr@ece.drexel.edu

ABSTRACT

Examining the bacterial or functional differences between multiple habitats/populations/phenotypes plays an important role in making inferences about the roles that the taxonomy and functional profiles can take on in microbial ecology. It is therefore important to the field of comparative metagenomics, using α- & β-diversity, that methods or algorithms can detect the importance of particular subsets of variables that best differentiate the multiple phenotypes in the data. Given todays genomic *data deluge* efficient methods that can carry out these inferences cannot be understated enough. We assume observations are collected from a multitude of different environments (e.g., males vs. females, control vs. stimulus, etc.), and each observation is comprised of hundreds or thousands of different taxa/functional features (i.e., 16S or whole genome shotgun). Our goal in this work is to examine the role, assumptions, and inferences that feature subset selection can provide the field of microbial ecology and comparative metagenomics. Specifically we examine feature subset selection algorithms using embedded and filter approaches to infer taxa importance on data collected from the human gut microbiome We compare several widely adopted approaches from machine learning including greedy algorithms and l_1 regularization methods, as well as some software tools provided with QIIME, on data collected from the American Gut Project and other canonical studies of the human gut microbiome. We find that there are very few OTUs that carry information in regards to predicting the sex of a gut sample, and that *Bacteroidetes* is quite frequently found in the top ranked OTUs.

1. INTRODUCTION

The amount of data being generated by fields in life science is almost unfathomable to what what was being generated just a decade ago. The field of genomics has observed data growth rates as well [20], and the cost of collecting such data is more rapid than Moore's law could have predicted[1]. Given this *data deluge* in genomics, it is apparent to the success of the field that comparative analysis tools of today are capable of scaling to the data of tomorrow, and still give researchers meaningful interpretations of their data. The engineering and computational intelligence community has already begun to perform translational research that can benefit the many fields within the life sciences [27], and specifically, the field of machine learning can provide extremely useful information to researchers in the life sciences.

In this work, we specifically focus on areas related to comparative metegenomics (and 16S analysis). Specifically, we assume that the raw sequences from the environmental samples have already been classified into operational taxonomic units (OTUs), or functions. The raw OTU[2] counts are stored in a matrix $\mathbf{X} \in \mathbb{N}_+^{K \times N}$, where $\mathbb{N}_+$ is the set of positive natural numbers, K is the number of OTU clusters, and N is the number of samples collected. The N samples contain a significant amount of *metadata* describing the sample, which is were we obtain phenotypes describing the sample. While there may be quite a few pieces of metadata, we shall only focus on one piece of metadata at a time. For example, a sample may contain the sex, age, and height of the person from where a sample was collected, and the analysis would only use *one* of those fields. That is we could use $\mathbf{X}$ to build a predictive model of sex. Both the data matrix and metadata can be found for hundreds of datasets though pioneering projects such as MG-RAST [19], KBase [5], the Human Microbiome Project [24], and the Earth Microbiome Project [11].

The data matrix $\mathbf{X}$ is comprised of many OTUs (e.g., the data from the American Gut Project has 25k+), although, as we shall show, many of them may carry little information about differentiating something such as the sex of the sample. Therefore, we use feature subset selection algorithms to detect the number of relevant features, or OTUs, that best differentiate the phenotypes. The field of feature subset selection can provide these types of insights into the importance of OTUs, and recent efforts have begun to scale these methods to massive datasets to keep up with todays data needs [31, 17]. We also show how one recent subset selec-

*Corresponding author.

[1] Refer to http://www.genome.gov/sequencingcosts/

[2] We use OTU from this point forward. The analysis for working with functional profiles is identical in our work, therefore, we use OTUs for brevity.

tion method for big data can be useful for determining taxa importance via a statistical hypothesis test.

This manuscript is organized as follows: section 2 covers the background and related works to elements of feature subset selection and how it relates to works in comparative metagenomics. Section 3 describes the methods used to carry out the experiments in section 4. Finally, section 5 provides some concluding remarks on the results and directions for future research.

2. RELATED WORKS

The core of this work lies in that of supervised learning. That is learning a mapping from an input space $\mathcal{X}$ (e.g., OTUS) to an outcome (e.g., health status), which gives us a strong motivation to evaluate the features that best represent the problem prior to learning a predicting model, such as a SVM or k-NN classifier. Several recent works have begun to examine the role of supervised learning in the microbiome [14, 9], and multi-class learning [18]. In these works, feature selection has been considered from an embedded perspective, while our previous work sought to use an information-theoretic perspective [10, 6, 7]. In this work we continue to evaluate feature selection using embedded and information-theoretic methods, and we evaluate the amount of predictive information in the differentiation between male and females' OTU abundance profiles.

3. METHODS

Feature subset selection approaches typically fall one of three categories: *wrapper*, *embedded* and *filter* methods. Wrapper based algorithms use a score computed from cross validation performance of a chosen classifier to search the subspace of features that will yield a subset of features that minimize the score. Unfortunately these methods are overly complex and do not scale to big data even with "more efficient" implementations. Therefore, we do not examine wrapper based algorithms in this work.

Embedded methods jointly optimize the classifier's parameters and feature selector simultaneously [13, 12]. The difference between embedded and wrapper based approaches, is that the feature selection for embedded methods is built into the objective function being optimized, which is not the situation for a wrapper method. Filter-based methods decouple the feature selection objective from classification by scoring feature independently from an error function. Hence, filter are typically very fast compared to wrappers or embedded approaches. In this work, we examine the use of filter-based approaches using information theory and embedded methods that induced a sparse solution using a minimization of the l_1 norm of a parameter vector.

3.1 Greedy Algorithms, Information Theory & NPFS

One of the fundamental quantities in information theory that has been widely adopted for feature subset selection with filters is *mutual information*, which is given by:

$$\mathsf{I}(X;Y) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_{X,Y}(x,y)\log\frac{p_X(x)\,p_Y(y)}{p_{X,Y}(x,y)} \qquad (1)$$

where $p_X(x)$ is the marginal distribution over the random variable X and $p_{X,Y}(x,y)$ is the joint probability distribution over X and Y. Hence, mutual information is the scoring function for determining the set of features $\mathcal{F}$ that carry the most information about an outcome Y. A simple algorithm for feature selection with a filter is the *greedy forward selection search* which seeks to maximize feature scoring function $\mathcal{J}$, which is shown in Figure 1. The initializes the relevant feature set $\mathcal{F}$ to be empty. Then for k iterations an objective function $\mathcal{J}$ is maximized, and the feature that maximizes the expression is added to the relevant feature set, $\mathcal{F}$, and removed from the feature set, $\mathcal{X}$. Simply using mutual information as the objective function is a fast way for microbial ecologists to examine the relative importance of taxa in a study collected from environmental samples. Though simply using mutual information will not capture inter-feature dependencies. Using other objective functions, such as joint mutual information [29], captures some of the inter-feature dependencies.

```
Input: Collection of features $\mathcal{X} := \{X_i : i \in [K]\}$,
scoring function $\mathcal{J}$, and phenotype variables $Y$.
Initialize: $\mathcal{F} = \varnothing$
while $|\mathcal{F}| < k$ do
    • Compute next best feature
          $X^* = \arg\max_{X' \in \mathcal{X}} \mathcal{J}(X', Y, \mathcal{F})$    (2)
    • $\mathcal{F} \leftarrow \mathcal{F} \cup X^*$
    • $\mathcal{X} \leftarrow \mathcal{X} \backslash X^*$
end while
```

Figure 1: Pseudo code for search selecting features using a greedy algorithm that attempts to maximize $\mathcal{J}$.

Our recent work includes the development of the Neyman-Pearson Feature Selection (NPFS), which automatically detects the relevant features in a dataset using a generic scoring function [8]. Furthermore, NPFS is highly parallelizable, which allows it to be quite effective for very large datasets. NPFS works by mapping out random samples of the original dataset to a scoring function which makes a prediction on which features are relevant. All of the sub-datasets have the same number of features selected then in a reduction phase NPFS applied the Neyman-Pearson test to detect feature importance. In this setting, NPFS can detect the number of important OTUs simply by guessing k in Figure 1 for the scoring function and letting the hypothesis detect features that appear to be more important.

3.2 Regularization for Subset Selection

Section 3.1 presented a greedy algorithm and tools from information theory that can be used to select features that are deemed important by the scoring function. Now we present feature selection from an embedded perspective. Let $\mathbf{y}$ be a vector in $\{\pm 1\}^N$ containing a binary outcome (e.g., control or stimulus) and $\mathbf{X}$ be abundance matrix. Predictions are made on $\mathbf{y}$ with $\mathbf{X}^\mathsf{T}\theta$, where $\theta \in \mathbb{R}^K$. If many of the entries of θ were zero then we could view the inner product of θ

with $\mathbf{X}$ as a form of feature selection. To encourage sparsity in θ's solution, Tibshirani presented lasso, which adds a penalty to the l_1-norm of θ [25]. Formally, lasso is given by:

$$\theta^* = \arg\max_{\theta \in \Theta} \frac{1}{2N} \|\mathbf{y} - \mathbf{X}^{\mathsf{T}}\theta\|_2^2 + \lambda\|\theta\|_1 \quad (3)$$

where $\lambda > 0$, and $\|\cdot\|_1$ and $\|\cdot\|_2$ are the l_1- and l_2-norms, respectively. For lasso to be effective at feature selection, it is assumed that $K \gg N$, which is typically an acceptable assumption with 16S and metagenimic data because there are typically only a few samples and a large number of features. The elastic-net was developed to avoid lasso selecting all features when $K \gg N$ is not met (see [32]). The objective function of the elastic net is given by:

$$\theta^* = \arg\max_{\theta \in \Theta} \frac{1}{2N} \|\mathbf{y} - \mathbf{X}^{\mathsf{T}}\theta\|_2^2 + \lambda_1\|\theta\|_1 + \frac{\lambda_2}{2}\|\theta\|_2^2 \quad (4)$$

3.3 Normalization and Scaling $\mathbf{X}$

In this work we do not use the raw OTU abundance abundance rather we use a scaled version of the relative abundance, and Su et al. demonstrate that using normalizations can improve the interpretability of the analysis [23]; however, Su et al. did not evaluate normalization's impact on feature selection. We normalize the columns of $\mathbf{X}$ to produce a relative abundance matrix, $\mathbf{X}'$, which is then used for feature subset selection. We scale $\mathbf{X}'$ by a factor γ to avoid numerical issues.

4. EXPERIMENTAL RESULTS

We evaluated several publicly available software tools for feature subset selection – some of which are implemented for biological data formats (see †). Below is a summary of the algorithms tested:

- Fizzy†: Fizzy is a feature subset selection tool for biological data formats that is built on top of Brown et al.'s FEAST Toolbox [2]. The experiments in this section use mutual information maximization (MIM) [16], unless otherwise stated. MIM uses (1) as the objective function in Figure 1. The Fizzy libraries were modified to directly output the mutual information scores (see GitHub for more details.)
- NPFS†: NPFS detects feature importance given a base subset selection algorithm [8]. We use command line tool for NPFS and use MIM as the base subset selection algorithm with 500 bootstraps.
- Lasso: Lasso is implemented in the Scikit-Learn machine learning package for Python [21]. We set $\lambda = 1$ and the maximum number of iterations to 1000.
- Elastic-Net: The elastic net is implemented in the Scikit-Learn machine learning package for Python [21]. We give equal weight to the penalization of the l_1 and l_2 norms, and the maximum number of iterations to 1000.
- Random Forest†: The random forest was implemented using QIIME's `supervised_learning.py` function [3]. The ensemble is generated with 500 trees using the out-of-bag error estimate. The random forest can rank features based on the average drop in accuracy if the feature is omitted from the dataset.

We have released code and data required reproduce the tabular results and figures. This supplemental information is available at https://github.com/gditzler/BigLS2014-Code. Furthermore, the data from the American Gut Project was obtained from https://github.com/biocore/American-Gut, and Caporaso et al.'s microbiome study was obtained from the Earth Microbiome Project (study identifier 550) [4, 11]. Our primary motivation for selecting benchmark datasets collected from the human gut was because it has been well studied, and becoming increasingly more understood [4, 1, 15, 26, 22, 30]. Thus, we can refer back to existing literature to verify that our results go along with our intuition. The metadata values for the sex of the sample is used as the phenotype (i.e., class labels). The data from the American Gut Project contains 469 samples (231/238 male/females), and 25703 OTUs. The Caporaso data contains 467 samples (336/131 males/females) with 16703 OTUs.

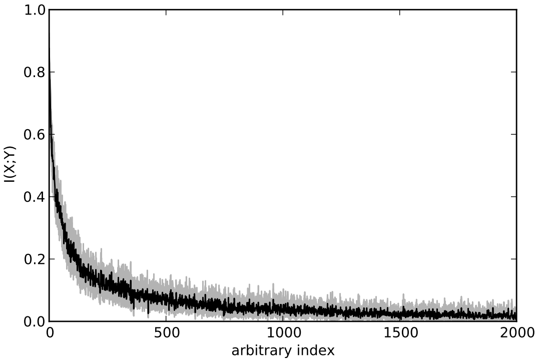


Figure 2: Mutual information shared between the taxa and sex phenotype for data collected from the American Gut Project. The mutual information is calculated over 50 bootstraps of the top ranked 2000 of 25k+ taxa.

4.1 On the Information in Taxonomy

In our first experiment, we evaluate the amount of mutual information that can be found in the OTU abundance table from the American Gut Project. To examine this, bootstrap samples are drawn from the entire data and the mutual information is computed. For clarity, only the top 2000 OTUs mutual information are reported in Figure 2. Note that we have sorted the x-axis according to the mutual information level. The grey line indicates the MI from one of the bootstrap trials and the black line is the average from mutual information.

One of the primary observations to make is that the vast majority of the OTUs have very little information shared with the sex phenotype. In fact, we can infer that there are only – approximately – OTUs that are somewhat informative for differentiating the sex of a sample. NPFS detects 73 OTUs as being important, and 30 of them are unique, which reconfirms that there are very few informative features for differentiating the sex of the sample.

Figure 3 shows another view of the mutual information in the American Gut Data; however, in this figure the mutual information is calculated as $\mathsf{I}(X_i; X_j)$, which ignores the phenotype. Again, we find relative few of the relevant features (as determined by $\mathsf{I}(X;Y)$), share information with other top ranked OTUs as determined by MIM. Similar observations can be made when lasso is evaluated on this data. Figure 4(a) shows the weights of lasso (θ) with the largest magnitude, and again the observation is that there are relatively few features (<100) that appear to be informative for predicting the sex phenotype. Figure 4(a) was generated with $\lambda = 1$. Figure 4(b) shows that there is relatively little variation in lasso's mean squared error (MSE) when evaluated as a function of λ. Elastic nets provide nearly identical results for data from American Gut Project, therefore, we have omitted the results to avoid redundancy.

4.2 Detecting Taxa Importance

The previous section solely focused on the examining the *general* amount of mutual information found in sex phenotype of the American Gut Project; however, we did not examine which OTUs are being ranked as important. Below is a list of the top 10 ranked OTUs as determined by the magnitude of the weight lasso assigns to each OTU.

- *Bacteria, Bacteroidetes, Bacteroidia, Bacteroidales, Bacteroidaceae, Bacteroides uniformis*
- *Bacteria, Bacteroidetes, Bacteroidia, Bacteroidales, Prevotellaceae, Prevotella copri*
- *Bacteria, Firmicutes, Clostridia, Clostridiales, Ruminococcaceae*
- *Bacteria, Firmicutes, Erysipelotrichi, Erysipelotrichales, Erysipelotrichaceae, Eubacterium dolichum*
- *Bacteria, Firmicutes, Clostridia, Clostridiales, Ruminococcaceae*
- *Bacteria, Bacteroidetes, Bacteroidia, Bacteroidales, Bacteroidaceae, Bacteroides*
- *Bacteria, Proteobacteria, Gammaproteobacteria, Cardiobacteriales*
- *Bacteria, Cyanobacteria, Chloroplast, Streptophyta*
- *Bacteria, Proteobacteria, Gammaproteobacteria, Xanthomonadales, Sinobacteraceae*
- *Bacteria, Firmicutes, Clostridia, Clostridiales, Lachnospiraceae, Blautia producta*

Furthermore, the top ten OTUs that NPFS detects as relevant are given by:

- *Bacteria, Bacteroidetes, Bacteroidia, Bacteroidales, Bacteroidaceae, Bacteroides*
- *Bacteria, Bacteroidetes, Bacteroidia, Bacteroidales, Bacteroidaceae, Bacteroides*
- *Bacteria, Firmicutes, Clostridia, Clostridiales, Lachnospiraceae, Blautia*
- *Bacteria, Proteobacteria, Betaproteobacteria, Burkholderiales, Alcaligenaceae, Sutterella*
- *Bacteria, Firmicutes, Clostridia, Clostridiales, Ruminococcaceae*
- *Bacteria, Bacteroidetes, Bacteroidia, Bacteroidales, Bacteroidaceae, Bacteroides*
- *Bacteria, Firmicutes, Clostridia, Clostridiales*
- *Bacteria, Firmicutes, Clostridia, Clostridiales, Ruminococcaceae*

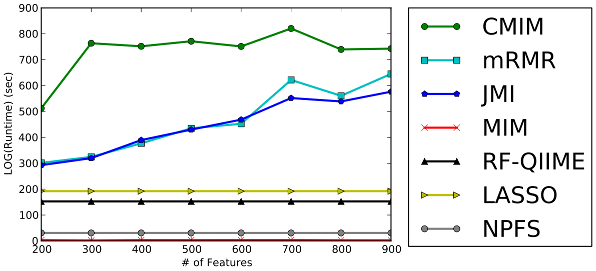


Figure 5: Runtimes seven algorithms that can be used for selecting important taxa. Note that times for NPFS, LASSO, and RF-QIIME are simply interpolated from 200 to 900 because they select a fixed number of taxa.

- *Bacteria, Firmicutes, Clostridia, Clostridiales, Ruminococcaceae, Ruminococcus*
- *Bacteria, Firmicutes, Clostridia, Clostridiales, Lachnospiraceae, Ruminococcus, gnavus*

where some OTUs are repeated because repeats in the 25k+ OTUs in the original dataset. For both lasso and NPFS, we observe *Bacteroidetes* is commonly being detected as OTU with high differentiation between the individual's sex, which *Bacteroidetes* has been hypothesized to be factor between male and female guts' differences [28].

4.3 Implementation Tradeoffs

Finally, we address the time it takes to evaluate several different subset selection algorithms on Caporaso et al.'s data collected from the Earth Microbiome Project. Figure 5 shows the runtimes of seven subset selection algorithms as a function of the number of features that the selection algorithm chooses (note that CMIM, mRMR, JMI and MIM are implemented in the Fizzy's feature selection). Its important to note that NPFS, LASSO and RF-QIIME can either: (a) detect the number of relevant features, or (b) weight the features and allow the user to choose what weight values qualify an OTU as being important. Hence the reason NPFS, LASSO and RF-QIIME is shown as being "fixed" w.r.t. the number of features being selected by the algorithm. From Figure 5, we observe that MIM has the fastest runtime, which is closely followed by NPFS. NPFS has the advantage of being extremely parallelizable, which our software implementation takes advantage of. The final runtimes for lasso, random forests, and NPFS were 192.2s, 152.7s, and 30.8s, respectively.

5. CONCLUSIONS

In this work, we evaluated embedded and filter based feature subset selection algorithms on data collected from the gut microbiome. We've shown there are very few OTUs that provide information for making these predictions using both information-theory and lasso. Furthermore, NPFS and MIM were shown to provide results from subset selection that go along with our intuition about the microbiome, while being extremely competitive in regards to runtime. NPFS's runtime can be improve further – nearly to the runtime of MIM – if more processor were available. This is because of NPFS's ability to be highly parallelized.

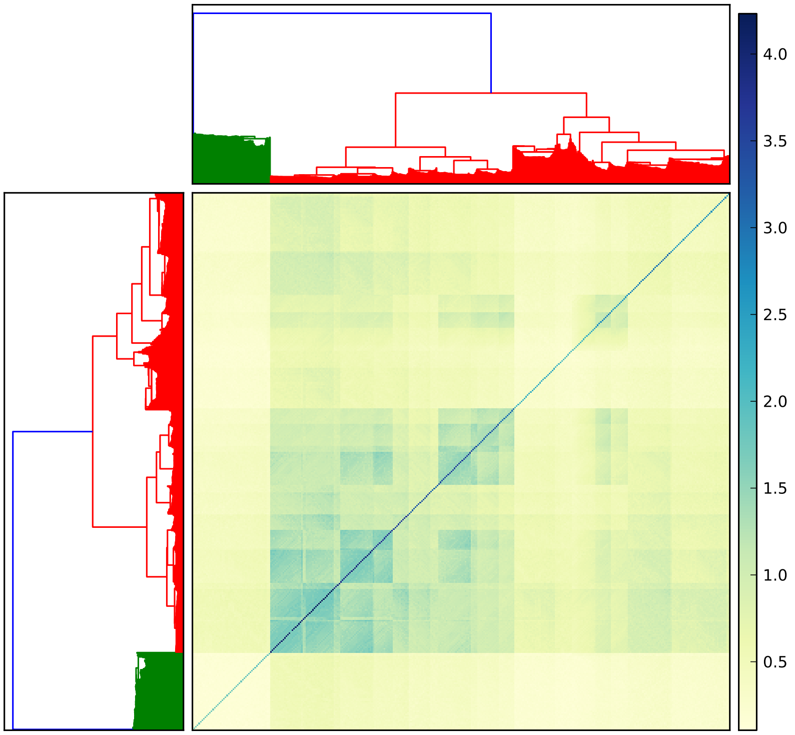
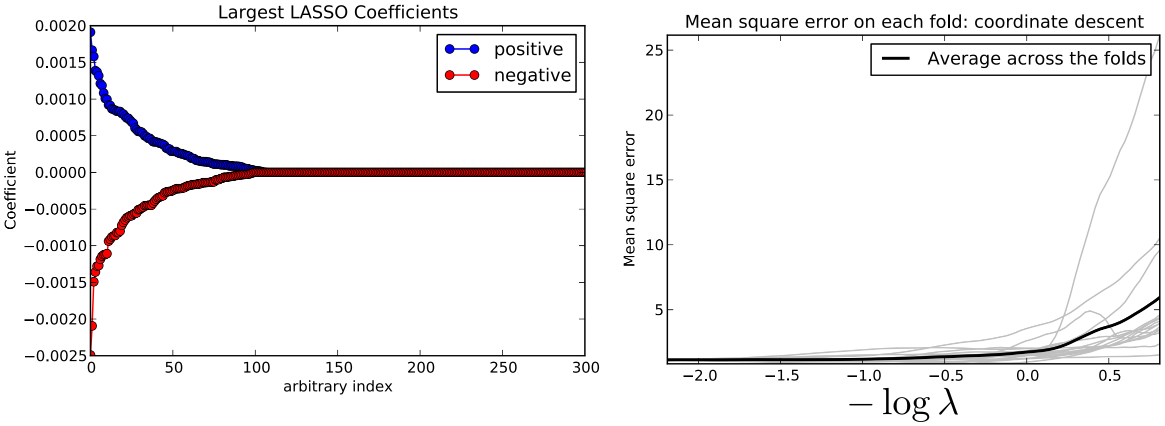


Figure 3: Mutual information, $\mathsf{I}(X_i;X_j)$, shared between two features X_i and X_j. The mutual information is reported in bits and only show the features with the largest mutual information calculated between X' and Y. Hierarchical clustering is performed with $D_{ij} = \frac{1}{\mathsf{I}(X_i;X_j)}$ (i.e., small value → high information and vice versa).



(a) lasso coefficients

(b) lasso MSE as a function of $\log\lambda$

Figure 4: Coefficients and the MSE of lasso applied to data from the American Gut Project.

6. ACKNOWLEDGEMENTS

G. Ditzler and G. Rosen are supported by the NSF CAREER award #0845827, NSF Award #1120622, and the Department of Energy Award #SC004335.

7. REFERENCES

[1] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, et al. Enterotypes of the human gut microbiome. *Nature*, 473:174–180, 2011.

[2] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, 2012.

[3] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7:335–336, 2010.

[4] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J. I. Gordon, and R. Knight. Moving pictures of the human microbiome. *Genome Biology*, 12(5), 2011.

[5] Department of Energy. DOE systems biology knowledge base, 2013.

[6] G. Ditzler, R. Polikar, and G. Rosen. Forensic identification using environmental samples. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1861–1864, 2012.

[7] G. Ditzler, R. Polikar, and G. Rosen. Information theoretic feature selection for high dimensional metagenomic data. In *International Workshop on Genomic Signal Processing and Statistics*, 2012.

[8] G. Ditzler, R. Polikar, and G. Rosen. A bootstrap based neyman–pearson test for identifying variable importance. *IEEE Transactions on Neural Networks and Learning Systems*, 2014.

[9] N. Fierer, C. L. Lauber, N. Zhou, D. McDonald, E. K. Costello, and R. Knight. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences*, 107(14):6477–6481, 2010.

[10] E. Garbarine, J. DePasquale, V. Gadia, R. Polikar, and G. Rosen. Information-theoretic approaches to SVM feature selection for metagenome read classification. *Computational Biology and Chemistry*, 35:199–209, 2011.

[11] J. Gilbert, F. Meyer, D. Antonopoulos, P. Balaji, C. T. Brown, C. Brown, N. Desai, J. A. Eisen, D. Evers, D. Field, W. Feng, D. Huson, J. Jansson, R. Knight, J. Knight, E. Kolker, K. Konstantindis, J. Kostka, N. Kyrpides, R. Mackelprang, A. McHardy, C. Quince, J. Raes, A. Sczyrba, A. Shade, and R. Stevens. Meeting Report: The Terabase Metagenomics Workshop and the Vision of an Earth Microbiome Project. *Standards in Genomic Sciences*, 3(3), 2010.

[12] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[13] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh. *Feature Extraction: Foundations and Applications*. Springer, 2006.

[14] D. Knights, E. K. Costello, and R. Knight. Supervised classification of human microbiota. *FEMS Microbiology Reviews*, 35(2):343–359, 2011.

[15] J. E. Koenig, A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh, R. Knight, L. T. Angenent, and R. E. Ley. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, pages 4578–4585, 2010.

[16] D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the Workshop on Speech and Natural Language*, pages 212–217, 1992.

[17] H. Liu, H. Motoda, R. Setiono, and Z. Zhao. Feature selection: An ever evolving frontier in data mining. In *Workshop on Feature Selection in Data Mining*, 2010.

[18] Z. Liu, W. Hsiao, B. Cantarel, E. F. Drábek, and C. Fraser-Liggett. Sparse distance based learning for simultaneous multiclass classification and feature selection of metagenomic data. *Oxford Bioinformatics*, 27(23), 2011.

[19] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(386), 2008.

[20] National Research Council. *Frontiers in Massive Data Analysis*. National Academies Press, 2013.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(2825–2830), 2011.

[22] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, M. Jian, Y. Zhou, Y. Li, X. Zhang, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, and S. D. Ehrlich. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464:59–65, 2010.

[23] C.-H. Su, T.-Y. Wang, M.-T. Hsu, F. C.-H. Weng, C.-Y. Kao, D. Wang, and H.-K. Tsai. The impact of normalization and phylogenetic information on estimating the distance for metagenomes. *IEEE Transactions on Computational Biology and Bioinformatics*, 2(9):619–628, 2012.

[24] The NIH HMP Working Group et al. The nih human

microbiome project. *Genome Research*, 19(12):2317–2323, 2009.

[25] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistics Society*, 58(1):267–288, 1996.

[26] P. Turnbaugh, M. Hamady, T. Yatsunenko, B. Cantarel, A. Duncan, R. Ley, M. Sogin, W. Jones, B. Roe, J. Affourtit, M. Egholm, B. Henrissat, A. Heath, R. Knight, and J. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 475:480–485, 2009.

[27] M. Vidyasagar. Opportunities in the life sciences. *IEEE Circuits and Systems Magazine*, 2012.

[28] P. Xu, M. Li, J. Zhang, and T. Zhang. Correlation of intestinal microbiota with overweight and obesity in kazakh school children. *BMC Microbiology*, 12(283), 2012.

[29] H. Yang and J. Moody. Data visualization and feature selection: New algorithms for non-Gaussian data. In *Advances in Neural Information Processing Systems*, 1999.

[30] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486:222–227, 2012.

[31] Y. Zhai, Y.-S. Ong, and I. W. Tsang. The Emerging "Big Dimensionality". *Computational Intelligence Magazine*, 9(3):14–26, August 2014.

[32] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67(2):301–320, 2005.