# Variable Selection to Improve Classification of Metagenomes

Gregory Ditzler,[1] Yemin Lan,[2] Jean-Luc Bouchot,[3] Gail Rosen[1]

[1]Department of Electrical & Computer Engineering, Drexel University, Philadelphia, PA 19104
[2]School of Biomedical Engineering, Science and Health, Drexel University, Philadelphia, PA 19104
[3]Department of Mathematics, Drexel University, Philadelphia, PA 19104

gregory.ditzler@gmail.com, yeminlan@gmail.com,
jean-luc.bouchot@drexel.edu, gailr@ece.drexel.edu

## 1 Introduction

Metagenomics is the study of DNA extracted from the microbial communities in an environment, in comparison to traditional genomics, which studies the nucleic acids from single organisms (Wooley et al., 2010). In a metagenomic study, a sample is collected directly from the environment, which can be a gram of soil (Rousk et al., 2010; Bowers et al., 2011), milliliter of ocean (Williamson et al., 2008), swab from an object (Caporaso et al., 2011), or a sample of the microbes associated with a host organism, such as humans (Caporaso et al., 2011; Costello et al., 2009). The microbial content of an environmental sample is termed its "microbiome". There are several questions that are of particular importance when the microbiome is being examined. In particular, *who is there*, *how much of each species is there*, and *what are they doing overall*? Some of these questions can be addressed using DNA/RNA sequencing followed by homology and taxonomic classification; however, usually hypotheses focus on answering: *which organisms and/or their functions (e.g., metabolisms) best differentiate multiple phenotypes in a collection of samples*? Consider a collection of gut microbiome samples that were collected from patients with *inflammatory bowel disease* (IBD), and a control set that do not have IBD. A natural question to ask when examining the differences between the gut microbiomes of the two phenotypes is what organisms or genes can distinguish patients with IBD and healthy controls? Knowing the answers to such question can be useful in developing a better understanding about a disease, and aid in developing medicines to target a disease cause.

The question of finding differentiating features, or variables of interest, has been deeply studied in the machine learning community (see Guyon et al. (2006) and Saeys et al. (2007)), which is commonly referred to as *feature selection*. Feature selection is the process of finding a subset of features that best differentiate between multiple classes, or in our case, phenotypes in a data set. The process of selecting features is typically achieved by maximizing some objective function (e.g., mutual information) in a greedy fashion. The central motivation for feature selection is to find a smaller subset of features that can be used to differentiate between the multiple phenotypes, which in turn can reduce the computational complexity of the classification algorithm tailored to do such a task. Furthermore, regression could be used instead of classification in the case of continuous-environmental variables; however, for this chapter, we assume that phenotypes takes on discrete states, and therefore, classification is the primary focus. Previously, feature selection has been shown useful to reduce the complexity of metagenome classification (Ditzler et al., 2012); however, in this article, it's use is expanded to determine relevance of biological features to associated phenotypes thus aiding researchers in drawing conclusions from metagenomic data.

Feature selection can be applied to a variety of metagenomic data (e.g., 16S rRNA, whole-genome shotgun, taxonomic annotations, gene annotations, etc.). In addition to selecting species which differentiate microbiomes, many studies wish to map DNA/RNA sequences to functional categories, and address enriched/depleted functions between samples. Depending on the type of question being asked and the nature of the data, there are a variety of functional databases to choose from. Table 1 highlights some of the most widely used databases. Large reference sequence database with a variety of functional descriptions are preferred because they provide detailed annotation of diverse dataset. This raw-labeling of sequences can provide much information, however, it cannot be used to analyze hierarchical functional structure in a dataset, such as *what high-level functions (e.g., reproduction/cellular transport) are upregulated in my sample?*. Instead, sequence labeling can answer *what genes exist in my sample?* or *which sample is functionally more diverse?*, because they provide better annotation coverage in the sample than higher-level databases. However, if it is required to annotate with well-defined vocabularies, which is needed to make biological inference and associations, then one wishes to use a standardized ontology database. For example, researchers can use Gene Ontology annotation to examine what functions are enriched in the sample compared to others. In some cases, researchers wish to annotate the function of a gene that appears in multiple organisms rather than just one. In other words, the focus is to accurately assign homologous genes associated with multiple species, which is especially important in metagenomics due to the complex mixture of organisms in a sample. Therefore, orthologous group databases are useful for annotating homologous function of orthologs. For studying a microbiome's metabolism rather than molecular functions, such as asking the questions – *what biological processes are enriched/missing from a diseased microbiome* or *should photosynthesis activity be enhanced in surface soil compared to deeper layer soil samples*, several metabolic pathway databases can be used. Finally, protein family databases search for conserved domains and motifs of protein sequences, and are important when considering the origin and evolution of proteins. For example, protein motifs that characterize pathogenicity may be used as potential targets for diagnosis and treatment.

Since the diversity of functional databases serves a variety of research questions, it is important to note that many studies would adopt several databases for annotation. Therefore, the optimal feature selection technique may depend on the database choice and the nature of taxonomic or functional data, such as the dimension of feature space, data sparsity, the possible range of fold-change between samples, *etc.*

Table 1: Functional databases mostly used for creating functional profiles.

| | |
|---|---|
| **Large collection of reference sequences** | |
| RefSeq | *Around 18 million proteins from 18k organisms, annotations are available for a subset of the database, well-annotated for human sequences.* |
| UniProtKB/Swiss-Prot | *Manually curated annotations for 500,000+ sequences, covering 12,930 organisms.* |
| **Standardized ontologies** | |
| Gene Ontology | *Well controlled vocabulary, primarily for eukaryotes.* |
| **Gene orthologous groups** | |
| COG | *Gene groups classified into 23 functional categories, inferred from 66 prokaryote and unicellular eukaryote genomes.* |
| KOG | *Eukaryote version of COG containing 7 eukaryotic genomes.* |
| eggNOG | *Automated annotation of orthologs in 1133 species.* |
| **Metabolism** | |
| KEGG Pathway | *400+ manually drawn pathways, based on reactions from multiple species.* |
| BioCyc/Metacyc | *2000+ single-organism, experimentally-derived pathways.* |
| SEED | *Subsystems that describe metabolic machinery with expert curation.* |
| **Protein domains and families** | |
| Pfam | *A large collection of protein families that share the same domain.* |
| FIGfam | *Protein families that share domains and pairwise align for their full length sequences, resulting in less sequences per family.* |

This chapter is organized as follows: section 2 highlights the components of a general feature selection algorithm and how to design such an algorithm. Section 3 presents the benchmark *MetaHit* data set, followed by an empirical analysis of feature selection algorithms tested on the *MetaHit* data set in section 4. Finally, section 5 draws concluding remarks for feature selection applied to metagenomic data.

# 2 Feature Selection

Feature selection can provide a unique insight about the variables that provide discriminating information about populations, or phenotypes, typically contained in the metadata. This metadata could be as simple as two populations, such as healthy or unhealthy, or significantly more complex by containing many different populations within a data sample. It is natural during the analysis of a biological data set to ask the question: *which variables provide the most differentiation between multiple populations*? The answer to such questions can be answered using feature selection (Guyon and Elisseeff, 2003).

There are several items to consider before applying a feature selection to a (biological) data set. First, how many features should be selected? Most feature selection algorithms assume that the end-user must select this parameter, and the quality of the results will most likely be highly dependent on the value of this parameter. In many situations, cross validation can be used to search for an acceptable value. Second, what is the primary objective for features selection? Is it the goal of the end-user to perform classification, or are they simply looking for the top $k$ features in the data set? The design of the objective function, $\mathcal{J}(\cdot)$, for feature selection can be used to emphasize, and address these questions.

Let $\mathcal{J}(\cdot)$ be a function of the features $X_j$ (for $j \in \{1, \dots, Q\}$), the label variables $Y$, and the current relevant feature set $\mathcal{F}$. Note that the collection of variables (e.g., operational taxonomic units, Pfams, *etc.*) is denoted by $\mathcal{X}$. The objective function can be designed in a way, such that it reflects the task at hand. For example, if a biologist is interested in the top ranking features that carry the most mutual information between $X_j$ and $Y$ then the objective function should reflect this goal. In this situation, using a mutual information maximization (MIM) method is sufficient to achieve this goal (Lewis, 1992). MIM can be implemented as follows: (a) compute $I(X_j; Y)$ for all $j$ ($I(X_j; Y)$ is the mutual information between $X_j$ and $Y$), (b) rank the mutual informations in descending order, (c) selection the top $k$ variables with the largest mutual information and place them in $\mathcal{F}$.

However, many times we seek to classify data based on $Y$, and in such situations designing a more complex objective function is required. For example, it may be more advantageous to select $\mathcal{F}$ in such a way that the features contained in $\mathcal{F}$ are informative about $Y$; however, they are not redundant (i.e., one or more features provide the same amount of information about $Y$). An example of such an objective function is given by

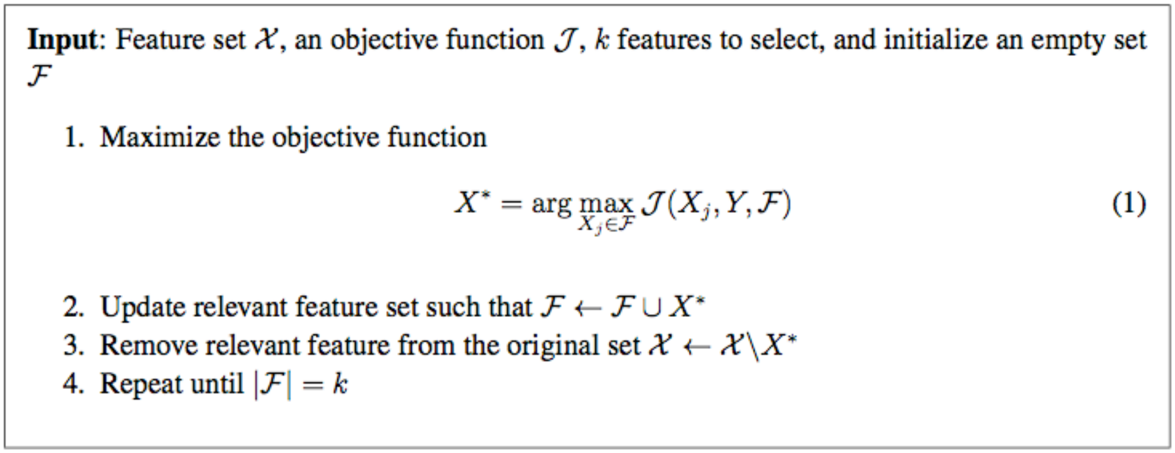
$$\mathcal{J}(X_j, Y, \mathcal{F}) = I(X_j; Y) - \sum_{X_s \in \mathcal{F}} I(X_j; X_s)$$

where the first term maximizes the mutual information between the features, $X_j$, and metadata $Y$, while the second term is penalizing $X_j$ for being redundant with the current relevant feature set in $\mathcal{F}$. The design of the objective function is quite important to the application to which feature selection is being applied. There are several works that highlight such results on bioinformatics data (Saeys et al., 2007), information theory methods (Brown et al., 2012), and general feature selection techniques (Guyon and Elisseeff, 2003).

A simple algorithm for feature selection is the *forward selection search*, which is shown in figure 1. The method begins by initializing the relevant feature set $\mathcal{F}$ to the empty set. Then for $k$ cycles equation (1) is maximized, and the feature that maximizes the expression is added to the relevant feature set, $\mathcal{F}$, and removed from the feature set, $\mathcal{X}$. The forward selection search is used with several feature selection objective function in section 4.

# 3 A Description of the MetaHit Database

As mentioned in section 1, feature selection can allow researchers in metagenomics to interpret the differentiating features in a data set. The interpretation can be insightful, and allow the researchers to determine the functional differences between multiple phenotypes. As a case study, let's examine a metagenome data set collected by Qin et al. (2010), which is widely referred to as the *MetaHit* data set. The data are collected from Illumina-based metagenomic sequencing of



```
Input: Feature set $\mathcal{X}$, an objective function $\mathcal{J}$, $k$ features to select, and initialize an empty set $\mathcal{F}$

  1. Maximize the objective function
       $X^* = \arg \max_{X_j \in \mathcal{F}} \mathcal{J}(X_j, Y, \mathcal{F})$    (1)
  2. Update relevant feature set such that $\mathcal{F} \leftarrow \mathcal{F} \cup X^*$
  3. Remove relevant feature from the original set $\mathcal{X} \leftarrow \mathcal{X} \backslash X^*$
  4. Repeat until $|\mathcal{F}| = k$
```

Figure 1: Generic forward feature selection algorithm for a filter-based method.

124 fecal samples of 124 European individuals from Spain and Denmark. The *MetaHit* data set represents one of the most comprehensive studies of the human gut microbiome. Among the 124 individuals in the database, 25 are from patients who have *inflammatory bowel disease* (IBD), and 42 patients are also obese. It is interesting to note that only three of the individuals who have IBD are also obese. Let us consider two different labeling schemes for the data: IBD and obesity, both of which are binary prediction problems. The sequences from each individual are functionally annotated using the Pfam database (Finn et al., 2010), in a recent study that utilized the *MetaHit* data set for feature selection on patient age (Lan et al., 2013). There are a total of 6,343 unique functional features detected in the data set, and figure 2 shows the $\log_{10}$ of the total abundance for each of the 6,343 functional features over the 124 observations in the data set.

One way to (loosely) access the separability of the IBD and no IBD patients (or obese and not obese) in the data, is to examine the *principal coordinate analysis* (PCoA) plots of the patients' Pfam data (Gower, 1967). Figure 3 shows the PCoA scatter plots of the two sample labeling schemes using PCoA implemented with the Euclidean distance. From these plots we observe that there is a significant amount of overlap between the classes for both labeling schemes.

# 4 Data Analysis

In this section, the classification accuracy and area under the receiver operating characteristic (auROC) curve for the the *MetaHit* data set are examined when feature selection is applied. The accuracy is measured using the standard 1–0 loss, and the auROC is interpreted as the probability of ranking a target data instance higher than a randomly selected non-target data instance (Fawcett, 2006). The IBD/Obese class label is identified as the target for the calculation of the auROC. The joint-mutual information feature selection algorithm (JMI) is implemented with a forward selection search, and the naïve Bayes classifier is implemented with a multinomial model. The FEAST feature selection toolbox implements the JMI algorithm (Brown et al., 2012). All statistics are presented as averages from 10-fold cross validation using stratified sampling. Stratified

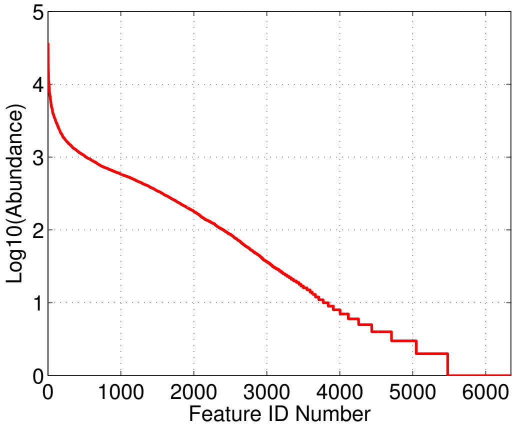
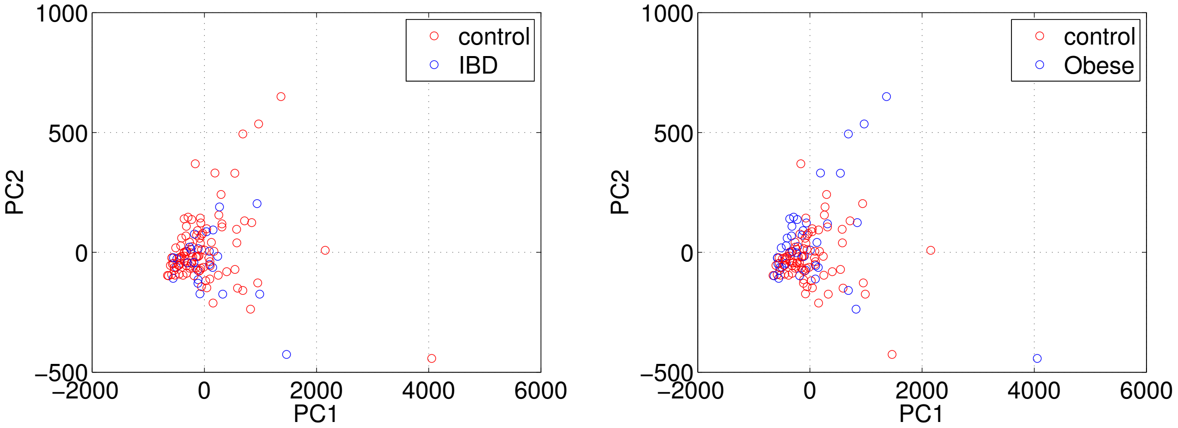


Figure 2: Logarithm of the total abundance of each feature detected by the Pfam database for Qin et al. (2010)'s human gut microboime data set. The $x$-axis represent rank of each feature corresponding with the number of detections sorted in descending order. From the plot, it is obvious that there are few Pfams with a large abundance and many Pfams with a very low abundance count. For example, there are 2,572 Pfams with 10 or fewer occurrences across the 124 observations.



(a) IBD

(b) Obese

Figure 3: Multidimensional scaling of the *MetaHit* data set with the IBD and Obese labeling of the samples. There appears to be a significant amount of overlap between the controls and targets for both prediction problems.

Table 2: Area under the ROC (auROC) curves and classification error for a naïve Bayes classifier tested using 10-fold cross validation.

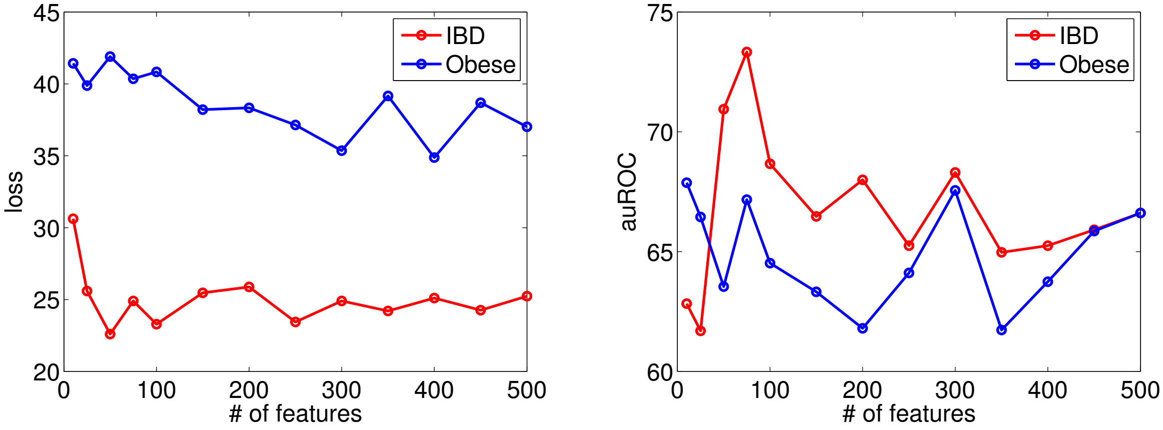
| | **auROC (IBD)** | **Error (IBD)** | **auROC (Obese)** | **Error (Obese)** |
|---|---|---|---|---|
| 10 | 0.706 | 0.233 | 0.640 | 0.395 |
| 15 | 0.624 | 0.290 | 0.672 | 0.352 |
| 25 | 0.616 | 0.292 | 0.660 | 0.403 |
| 50 | 0.750 | 0.223 | 0.649 | 0.422 |
| 100 | 0.660 | 0.249 | 0.659 | 0.397 |
| 200 | 0.654 | 0.257 | 0.643 | 0.389 |
| 500 | 0.635 | 0.277 | 0.641 | 0.378 |
| All | 0.665 | 0.238 | 0.622 | 0.240 |

sampling assures that instances from each class will be in each cross-validation data set. Note that completely random cross-validation data set partitions do not guarantee this property.

The auROC and loss for the multinomial naïve Bayes classifier are measured using the two labeling schemes described in section 3 (i.e., IBD and obese). Table 2 contains the classification assessments from the different labeling schemes as well as a variation in the number of features that are selected via JMI. From table 2, it is clear that feature selection can have a significant outcome in the classification results. This is best shown in figure 4 which shows the number of features selected by the MIM algorithm vs. the loss (figure 4(b)), and the auROC (figure 4(a)). Note that these results are generated using the mutual information maximization approach; however, similar results/trends are observed for other feature selection methods.

Figure 5(a) presents a visualization of the *MetaHit* data set before and after MIM feature selection is applied. The features are sorted from high to low in terms of overall abundance, and the patients are represented such that samples 1 through 99 do not have IBD, and samples 99 through 124 have IBD. Clearly, this shows a large amount of sparsity that is inherent in the data, which would also be evident if taxonomic abundances were used over Pfams. Figure 5(b) shows that most of the features being selected by MIM are relatively abundant features; however, simply because a feature is abundant does not imply that the feature is relevant. This can be observed near the 44th feature in figure 5(b). Note that the features in figure 5(b) are order by the time the were selected by the forward search.

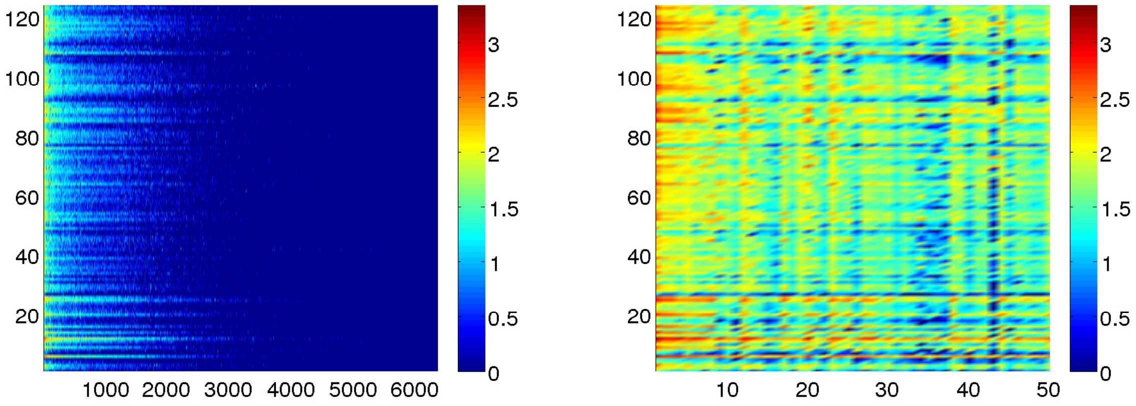
The top Pfams that maximize the mutual information for the *MetaHit* data set are shown in table 3. It is known in IBD patients, the expression of ABC transporter protein (PF00005, the first feature MIM selected for classifying IBD vs. no IBD samples) is decreased which limits the protection against various luminal threats (Deuring et al., 2011). The feature selection for IBD also identified glycosyl transferase (PF00535), whose alternation is hypothesized to result in recruitment of bacteria to the gut mucosa and increased inflammation (Campbell et al., 2001). And the genotype of acetyltransferase (PF00583) plays an important role in the pathogenesis of IBD, which is useful in the diagnostics and treatment of IBD (Baranska et al., 2011). It is not surprising that ABC transporter (PF00005) is also selected for obesity, which is known to mediate fatty acid



(a) loss of naïve Bayes

(b) auROC of naïve Bayes

Figure 4: The effect of the number of features selected by the MIM algorithm vs. the loss (*left*), and the auROC (*right*). The number of features being selected has a larger effect on the auROC (i.e., detection of target population examples), than the accuracy of the system. Similar results are observed with JMI and other feature selection methods.



(a) no feature selection

(b) feature selection

Figure 5: Visualization of the abundance matrix (on a $\log_{10}$ scale) (a) before and, (b) after MIM feature selection. The $x$-axis represents a feature and $y$-axis represents samples. Sample 1 through 99 do not have IBD, and samples 99 through 124 have IBD. (b) contains the top-50 features relevant to the 124 datasets. Differences between the two classes cannot be visualized, however, classification auROCs are 10-15% above chance.

transport that is associated with obesity and insulin resistant states (Ashrafi, 2007), and ATPases (PF02518) that catalyze dephosphorylation reactions to release energy.

Table 3: List of the "top" Pfams as selected by the MIM feature selection algorithm. Note that redundancy terms are not accounted for in the objective of MIM. Hence the feature below are the ones that provide the largest amounts of mutual information. The ID in parenthesis is the Pfam accession humber.

| | **IBD features** | **Obese features** |
|---|---|---|
| *feature 1* | ABC transporter (PF00005) | ABC transporter (PF00005) |
| *feature 2* | Phage integrase family (PF00589) | MatE (PF01554) |
| *feature 3* | Glycosyl transferase family 2 (PF00535) | TonB dependent receptor (PF00593) |
| *feature 4* | Acetyltransferase (GNAT) family (PF00583) | Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase (PF02518) |
| *feature 5* | Helix-turn-helix (PF01381) | Response regulator receiver domain (PF00072) |

# 5 Conclusion

This chapter has presented a broad overview about how feature selection algorithms can be used to facilitate and interpret data in the field of metagenomics. Recall that metagenomic abundance data can be of very large dimension (e.g., *MetaHit*), and feature selection reduces the dimensionality of the space to allow for a quick interoperation of the data. Furthermore, feature selection is also useful for classification because it allows us to remove potentially irrelevant features from the data set, which allows the classier to focus on learning from the relevant information rather than attempt to decipher what is or is not relevant.

# References

Ashrafi, K. (2007). Obesity and the regulation of fat metabolism.

Baranska, M., Trzcinski, R., Dziki, A., Rychlik-Sych, M., Dudarewicz, M., and Skretkowicz, J. (2011). The role of n-acetyltransferase 2 polymorphism in the etiopathogenesis of inflammatory bowel disease. 56:2073–80–.

Bowers, R. M., McLetchie, S., Knight, R., and Fierer, N. (2011). Spatial variability in airborne bacterial communities across land-use types and their relationship to the bacterial communities of potential source environments. *ISME Journal*, 5:601–612.

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

Campbell, B. J., Yu, L. G., and Rhodes, J. M. (2001). Altered glycosylation in inflammatory bowel disease: a possible role in cancer development. 18:851–8–.

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

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science*, 326:1694–1697.

Deuring, J. J., Peppelenbosch, M. P., Kuipers, E. J., van der Woude, C. J., and de Haar, C. (2011). Impeded protein folding and function in active inflammatory bowel disease. 39:1107–11–.

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

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunesekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S., and Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Research*, 38:D211–222.

Gower, J. (1967). Multivariate analysis and multidimensional geometry. *Journal of Royal Statistics Society*, 17(1):13–28.

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

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

Lan, Y., Kriete, A., and Rosen, G. L. (2013). Selecting age-related functional characteristics in the human gut microbiome. 1:–.

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

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

Rousk, J., Bååth, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., Knight, R., and Fierer, N. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME Journal*, 4:1340–1351.

Saeys, Y., Inza, I., and Larra naga, P. (2007). A review of feature selection techniques in bioinformatics. *Oxford Bioinformatics*, 23(19):2507–2517.

Williamson, S., Rusch, D., Yooseph, S., Halpern, A., Heidelberg, K., Glass, J., Andrews-Pfannkoch, C., Fadrosh, D., Miller, C., Sutton, G., Frazier, M., and Venter, J. C. (2008). The Sorcerer II global ocean sampling expedition: Metagenomic characterization of viruses within aquatic microbial samples. *PLoS Biology*, 3(1).

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology*, 6(2):1–13.