The Earth Microbiome Project from a Data Science Perspective

Jean-Luc Bouchot, Gregory Ditzler, Gail L. Rosen

Abstract

The Earth Microbiome Project (EMP) is a repository for microbiome studies which are carried out by researchers around the globe. Together with the genomic content of their samples, researchers have released information about the state of the host or habitat, which serves as the sample metadata. These metadata include, but are not limited to, salinity, pH, or the sex of the host. We analyze this dataset from a data science perspective.

After describing the global aspects of the EMP we investigate two problems using information-theoretic tools for feature selection. We first describe an algorithm to find which taxa found in a human gut are most related to the sex of a host. Secondly we investigate the problem of predicting the class of shrimps in a frozen ocean sample. We see that analyzing the abundance of a very small subset of the taxa enables the prediction.

1 Introduction

1.1 The Earth Microbiome Project

Metagenomics is concerned with the study of genomic data of microbial communities in an environment. The microbial population extracted from an environmental sample is denoted as its microbiome. Recent advances in deep-sequencing technologies have made it possible to collect many samples from various environments. Among other projects, the Earth Microbiome Project (EMP)\(^1\) appears as one of the challenging platforms in metagenomic data analysis. As of October 1\(^{st}\) 2013 the EMP contains more than 100 different studies coming from different environments: soil, fecal, ocean, and human. All of these studies contain thousands of different samples that are composed of thousands of sequences. The EMP represents a total of 20,000+ samples that may be used for further analysis of microbial communities. Not only do all the studies record DNA (or genetic) information but most of them have a rather comprehensive description of the external environmental parameters. The meta-parameters measured are diverse, ranging from very general information (such as the machine used for the deep-sequencing or which center is responsible for a given study) to more specific information (such as the salinity, elevation, altitude, etc... ).

The aim of this paper is two-fold. We first give a rather detailed overview of the EMP. Of particular interest is its usability from a data scientist’s perspective. As detailed later, the EMP has a lot meta-data; however, there is a significant amount of the meta-data missing between studies because each study was carried out by different groups trying to examine one component of the microbiome. For example, samples collected from a swap of saliva may not have had a

\(^1\)http://www.microbio.me/emp/
pH measurement collected; however, another study sampling data from soil may have the pH measurement. Thus, the pH measurement is missing from all the samples collected from the saliva because the measurement was never recorded. By releasing our tools and results, we hope to help researchers find their way through the EMP datasets and ease their decision in picking relevant studies that serve their interests. Due to the high variability of the different studies, first understanding the different studies and the EMP is mandatory before trying information retrieval.

As a second contribution, we describe some basic information-theoretic tools for analyzing EMP studies. In particular, we try to find relevant operational taxonomic units (OTUs, corresponding to microbes that classified together within a given similarity – 97% for the EMP) that are able to carry information about environmental parameters. We show that by doing this, we are able to drastically reduce the size of the data we are dealing with. This second part is important to biologists at least for two reasons: 1) the set of parameters measured are not necessarily consistent across studies (e.g. the parameter TEMP representing the temperature of the subject is measured in only 24 different studies – and has some entries missing in these studies – but not in the remaining 77) or 2) some data have been measured only for a subset of samples in a study (see for instance parameter SHRIMP that subclassifies the shrimps analyzed in study 6782 – someone might want to classify the remaining unknown samples). In both cases, one wants to complete the data, either using only information within a given study or across different, but similar, studies.

One particular aspect of the EMP is that it is a collaboration of many different labs. The studies are therefore quite different from one another and having a consistent analysis across studies is almost unrealistic. We introduce some useful tools to analyze such databases and facilitate its understanding by introducing methods to extract relevant features from the soup.

1.2 Motivations

Due to recent advances in sequencing technologies we now have access to very large dataset containing genomic information. The challenges are multiple. How can we use the data from a study to infer some information about another study? How can we use the information within a study and across multiple studies to complete missing measurements? In terms of classification, can we find common bacteria in different studies that share similar features or similar environmental conditions? These are examples of questions that need to be answered.

However, due to the nonuniformity of the sampling process between different studies, it is difficult to start looking at similarities. Do we want to compare all studies together? Wouldn’t it be preferable to use only ocean sample to complete the salinity parameter? We try here to answer some of these questions in some particular cases.

1.3 Organization - Structure

This paper is organized as follows. In Section 2 we introduce the primary tool used to analyze and extract statistics out of a study. In particular we describe its features and how to use it. Section 3 presents a first a general then a more exhaustive description of the EMP datasets as of October 1st 2013. Section 4 introduces information-theoretic tools for relevant feature extraction.

\(^2\)Laverock ocean acidification
2 Preprocessing EMP data

2.1 Presentation of the software tools

Our software tools\textsuperscript{3} are intended to make the analysis of the EMP studies easy, while requiring few other software dependencies. Furthermore, the software is available with an open-source license. An online page is kept updated with some general statistics as well as some useful routines to extract particular information out of the database. Examples of information obtained with these functions will be detailed in Section 3.

The second point is the easiness of its use. Indeed, software exists to work with the EMP dataset \cite{CKS10}, but require a cumbersome installation process. Our tools allow to open, extract, and export any study to a certain given format. This allows anyone to access a massive real-world dataset and transport it in their favorite software environment.

2.2 Usability

The website accompanying this article contains a set of bash files and Python scripts and modules. The data analysis module works using three files:

- **a biom file**: It contains the taxonomic information, sample IDs, taxa IDs, and some information about the file itself. The taxonomic information can be seen as the matrix of abundances, i.e. a matrix $A \in \mathbb{R}^{m \times N}$

- **a mapping file**: It contains all the important information to the data preprocessing as well as what we refer to the environmental parameters.

- **a features of interest file**: It specifies which metadata one is interested in as well as which format it is in. It is written in a tab-separated text file similar to:

\begin{verbatim}
SEX D T
AGE_IN_YEARS D I
PH C F
\end{verbatim}

where the first column corresponds to a parameter name, the second one to whether the parameter is a continuous or discrete variable, and the last column to the format (Integer, Text, Float).

2.3 Description of the main file

One of the first aspects of our tool is to let the researcher convert the data easily into their preferred format. For this we provide the Python function `output_missing_data_matrix.py` that generates four files: one corresponds to the biom file, another one to the subset of the mapping files containing only the interesting features, and the third one contains a dictionary translating the text parameters in integer class labels (for instance `MALE->0 FEMALE->1`). These four files are generated in the specified format.

\textsuperscript{3}The software tools and detailed analysis of the EMP are available and regularly updated from the authors’ website http://www.math.drexel.edu/~jb3455/emp-data.html
3 EMP and missing data

To understand the problem of working with missing metadata it is interesting to look at Fig. 1(a). It represents the whole EMP as single matrix. Each single row corresponds to a parameter taken from the 665 that exist in the aggregation of the studies. All the samples from the studies are concatenated such that the $n_1$ first columns correspond to the $n_1$ samples of study $S_1$, the $n_2$ following ones to the samples from study $S_2$ and so on until the 101$^{st}$ study (as of October 1$^{st}$ 2013). This yields a binary matrix of about 665 rows by more than 20$k$ columns where a 1 (light pixel) represents where there is a missing entry. It is clear that most of the data are missing; this is mainly due to the fact that not all the parameters are collected across all studies – for instance, there is no need in a SEX parameter for a soil sample. What is even more interesting is if we remove some useless information (such as some extensive word descriptions of the samples, see for instance parameters Description, or Duplicate) or some redundant information (such as the SAMPLE_CENTER or STUDY_CENTER), then we are left with the matrix illustrated on Fig. 1(b). This reduced matrix contains 372 rows with a high proportion of missing data and most of the black lines have disappeared implying that most of the data accessible are actually useless from a data scientist’s perspective.

![Binary visualization of the missing entries of the EMP](image1)

(a) Binary visualization of the missing entries of the EMP

![Missing entries of a relevant subset of the parameters of the EMP](image2)

(b) Missing entries of a relevant subset of the parameters of the EMP

Figure 1: Visualization of the missing entries of the environmental parameters of the Earth Microbiome Project (white pixels represent a missing entries while black is an entry filled). Each row corresponds to a given parameter and each column to a given sample.
3.1 Global aspects of the Earth Microbiome Project

To get a better understanding of the challenges associated with the EMP we give some global statistics on it. As mentioned in the previous Section, the EMP dataset is comprised of 101 studies. This represents a total of about 20,000 samples most of which are taken from 8 different countries around the world (all the continents are represented - unevenly). Altogether, the EMP gathers over 1,700,000,000 sequences. When merged together, we have access to a grand total of around 30,000 unique OTUs. Not every sample contains all of the different OTUs and in general only few of them will actually be present. Biological facts aside, the EMP distinguishes itself by adding environmental parameters. Some of the physicals have been measured or controlled for the different samples. As a consequence, a total of more than 650 different external parameters have been noted at least once. However many parameters have been measured only for a particular study.

3.2 Looking at the studies

The studies differ from one another in the number of samples and environment they come from. The main habitats are human, animal, lakes, and soil. Even within a habitat the diversity is large (we give an example in the last section). For instance a human microbiome sample may be taken from the skin, the mouth, or the gut. Therefore it is very important before starting any analysis on the EMP to find the relevant studies to our problems.

Each study contains an average of 195 samples and contains an average of 5,958 unique OTUs. The vector of OTUs for a sample embedded in the 30,000 known in the dataset is very sparse. Some of the parameters are study-dependent and not sample-dependent (e.g. sequencing center, sequencing machine, sample IDs, PI of the project, ...). If we decide to consider only sample-dependent metadata we are left with an average of 20 environmental metadata parameters. While this amount of information is appealing to scientists, it becomes challenging to infer information because many of them are missing: on average a whole study contains more than 700 missing metadata. Many times these parameters have not been measured at all within a study. Given a study and considering the metadata parameters that are missing at least once, on average, less than 50% of the samples have the metadata measured. This fact makes it very difficult to derive stable and robust machine learning algorithms. It also justifies a future interest as how to combine different studies to access more information. Another interesting observation with merging several studies in the EMP is that we have an average of 125 samples are not complete per study. This represents more than half of the samples.

4 Using information theory for feature selection and data understanding

We describe here a process for feature selection in metagenomic studies. Feature selection techniques become mandatory in bioinformatics due to the large dimensionality of the objects (here in the sense of mathematical vectors containing the OTU abundances) with which we deal [SIL07]. Recently, such ideas have been successfully applied to several areas of biology such as predicting the age of gut microbiomes [LKR13], processing mass spectrometry data [GTABL13], finding
new drug molecules \([KMT^{+11}]\). We aim to offer some advice for the design of feature extraction algorithms \([DLBR13]\) for the EMP.

We first look at the study \(550^{\text{iv}}[CLC^{+11}]\) and try to find the relevant OTUs explaining the \textsc{sex} environmental parameter. In this work, we examine the analysis of the data from a biologist’s perspective and from a machine learning perspective. In the machine learning perspective, our problem is tailored towards data prediction or completion. In this case, we need to find a subset of biological features that carry enough explanatory power about the environmental parameter without redundancy. In the biological case, scientists are interested in finding functional or taxonomic dependencies or correlations (if not causalities). When it comes to finding important OTUs, detecting redundancies is not a problem and might actually be of interest in some cases.

In this work, we examine the use of greedy feature selection algorithm (see Algorithm 1) to select a subset of taxonomic units that best categorize the differences between the environmental parameters. Whether we are interested in the biological problem or in the prediction problem is reflected in the design of the objective function defined in Equation (1). Indeed, this objective function controls the way we select a feature based on how much information it brings to explaining an external variable. We justify here the use of information-theoretic cost functions that are suitable for EMP data \([BPZL12]\). For the biological problem (i.e. finding a subset of explanatory OTUs) we have

\[
J_{MI}(X, Y; \mathcal{A}) := \sum_x \sum_y p(xy) \log \frac{p(xy)}{p(x)p(y)} = I(X, Y). \tag{1}
\]

where a feature \(X\) (an OTU) and data \(Y\) (e.g. sex of the host) are represented by random variables. \(p(x), p(y), \text{ and } p(xy)\) represent respectively the probability distribution of feature \(X\), data \(Y\), and joint probability distribution of \(XY\). This function is known as the Mutual Information. In other words, the algorithm tries to maximize the information shared by a feature \(X\), and a data variable \(Y\). As we can see, this information criteria does not consider the active set of selected OTUs \(\mathcal{A}\).

For the data prediction problem we use the following criteria:

\[
J_{JMI}(X, Y; \mathcal{A}) := \sum_{G \in \mathcal{A}} I(XG, Y) = I(X, Y) - \frac{1}{|\mathcal{A}|} \sum_{G \in \mathcal{A}} [I(X, G) - I(X, G|Y)], \tag{2}
\]

where \(I(X, G|Y) = \sum_y p(y) \sum_{x,g} p(xy|z) \log \left( \frac{p(xy|z)}{p(x|z)p(y|z)} \right)\) represents the conditional mutual information. The information-theoretic function (2) is known as the Joint Mutual Information (JMI). In this case, not only do we want to increase the explanation potential of our active feature set, but we also add a constraint on the redundancy between two selected features. Indeed, a given feature will be selected if it shares enough information with the output, or metadata variable, without being redundant with the features already selected (i.e. present in the active set \(\mathcal{F}\)).

\(^4\)Entitled \textit{Caporaso Illumina Time Series}:
ftp://thebeast.colorado.edu/pub/QIIME_DB_Public_Studies/study\_550\_split\_library\_seqs\_and\_mapping.tgz
**Data:** Objective function $J$, a set of OTUs $F$, a vector of environmental variables $y$, a data matrix $X$, a number of features to select $k$

**Result:** A set of active (selected) OTUs $A$

\[
A \leftarrow \emptyset; \\
i \leftarrow 0; \\
\textbf{while } i < k \textbf{ do} \\
\quad \text{Maximize the objective function:} \\
\quad \quad X^* = \arg \max_{j: X_j \in F} J(X_j, Y; A); \\
\quad \quad A \leftarrow A \cup \{X^*\}; \\
\quad \quad F \leftarrow F \setminus \{X^*\}; \\
\quad \quad i = i + 1; \\
\textbf{end}
\]

**Algorithm 1:** Forward selection algorithm

4.1 Finding relevant OTUs Using Information Theory

As a use-case we consider the study 550 from the EMP which contains a total of 1,967 samples that are represented by 16703 OTUs. These samples were taken from one of the four different sites from human hosts: left palm (499 samples), right palm (493 samples), feces (467 samples), or mouth (508 samples). Using this study of bacterial samples collected from different body sites and individuals, we explore the use of *Mutual Information Maximization* (MIM) to select the 100 most informative bacteria. In the experiment, we use only the gut samples collected from the male and female subjects, and we use the bacterial abundances to differentiate the sex of the host. We then examine the abundance of the family and genus levels that were detected in the top 100 OTUs as ranked using MIM (see Figure 2). We find that the most abundant families are Ruminococcaceae, Lachnospiraceae, and Bacteroidaceae. There were fewer OTUs that were classified down to the genus level; however, we find that Bacteroides are the most abundant genus that was classified.

4.2 Data prediction

4.2.1 Controlled tests

In this case, we are trying to reduce the number of features that are needed to predict the output of a variable. Therefore, we want to make sure that the feature subset is small but still contains a sufficient amount of information to predict on the environmental variables. This is why we want to enforce the nonredundancy in the relevant feature set. We want to find OTUs that are able to predict the sex of a sample’s host via a *naïve bayes classifier* (NBC). It should be noted that using other classifiers will give different results, but the NBC gives good results with low complexity and training time. Also it is worth mentioning that the NBC is well known in the biological community for taxonomic classification for its robustness [RRR11].

In this section, all the tests are carried over in a similar way. 20 random partitions of the samples are created and each partition is divided into 10 non-overlapping subsets of the samples. The ratio of male and female hosts is preserved along the partitions. Then for each random partition we proceed to a cross-validation like scheme, where one of the subsets is used to measure
Figure 2: Histograms indicating the unique families/genus detected in the top 100 OTUs. Only the OTUs classified down to the specified level are considered out of the original 100 OTUs.

the accuracy of the NBC based on the OTUs selected. For these tests, we vary the number of features used and calculate the average, minimum, and maximum consistency of the active OTUs as well as the error on the test set.

The consistency index between two sets $\mathcal{A}, \mathcal{B}$, subsets of $\mathcal{F}$ is a measure of how the different feature sets overlap, within the given larger one. It is defined by Kuncheva [Kun07] as

$$C_F(\mathcal{A}, \mathcal{B}) := \frac{rK - k^2}{k(K - k)}$$

where $|\mathcal{A}| = |\mathcal{B}| = k$, $|\mathcal{A} \cap \mathcal{B}| = r$ and $|\mathcal{F}| = K$. The consistency has a value between $-k(K - k)$, when the subsets $\mathcal{A}$ and $\mathcal{B}$ do not have a single feature in common, and 1, achieved if and only if the two subsets $\mathcal{A}$ and $\mathcal{B}$ are the same.

Kuncheva’s index can be used as a criteria on the quality of the feature selection process: the closer it is to one, the more stable the feature selection is. Figure 3 shows the behavior of the consistency index, when considering only oral samples. From the Caporaso et al. study, we calculated the consistency between any two feature subsets within a partition and recorded their min, max, and average values. The results have been aggregated together for the 20 different random runs.

As we can see the index becomes stable as the size of the subset increased, until a plateau is reached (in the JMI case). MIM is stable from the beginning. This can be explained by the fact that the algorithm will constantly try to get the most information - independently from what has been picked before. The stability of the MIM feature set selection is not guaranteed once the most important features have been chosen. Then Kuncheva’s index drops. This could be explained by the fact that the OTUs most common to all samples have already been taken and that the algorithm start picking more rare OTUs. The index drops similarly in Figure 4.

The behavior of the first graphs from Figures 3 and 4 comes from different OTUs having different importance in characterizing the sex of the host. In other words we accumulate a major part of the information about the sex of the host with the first few features. The following add only little to no information to the problem. The OTUs selected first are mandatory in explaining
the environmental parameter as can be seen Kuncheva’s index that stagnates once about 100% are included. The accuracy is calculate as the $\ell_1$ norm of the difference of the prediction with the exact class. It corresponds as the total number of misclassified samples. Looking at Figure 5 we see that only once we reach a feature set of size $140 - 150$ we have accuracies that are reliable.

Recall that Figure 5 shows the accuracy when dealing with the oral subsets of the study 550 which contains around 508 samples. In other words, a maximum norm of 5 as reached for instance for the JMI objective function corresponds to a misclassification rate of less than 1%. Other misclassification results on the study 550 are found in Figure 6. In terms of performance, the gut samples seems to offer the best discrimination between sex as NBC achieves an average accuracy of less than 3 misclassified samples, corresponding to less than 0.65%. It is also worth mentioning that because the samples used in Figure 6(b) are from two different environment, it is harder for the algorithm and classifier to reach reliable results. However, this $\ell_1$ norm should now be seen on around 1000 samples.

Another interesting point is the amount of reduction reached with this feature extraction process. We started with an input sample in a 16,703-dimensional and was reduced to a vector

Figure 3: Evolution of the consistency index with respect to the feature set size with the oral samples.

Figure 4: Evolution of the consistency index with respect to the feature set size with the feces samples.
of around 100 dimensions (which corresponds here in a reduction of 99%) before any classification algorithm was applied.

### 4.2.2 Application

The point of all study is to help scientists understand and predict some unknown variables from very large amount of information. Consider the study 678 from the EMP [LST+10, LKT+13], it contains 276 samples represented by 7,785 OTUs. All samples come from the same environment and habitat, which simplifies the study, compared to the previous one. These 276 samples are divided into three categories, derived from the SHRIMP parameter:

- 40 Callianassa,
- 213 Upogebia,
- 23 unknown.

We try here to predict the potential shrimp-class of the unknown samples. However caution should be taken when considering the results: 1) The sample size is somewhat small - better
theoretical results can be achieved with a larger training set, 2) We have no way to verify our results, and our prediction would need to be assessed by a trained biologist. Moreover, the two classes are very imbalanced and learning a classifier in these conditions is always a tough task [HG09]. It is very important to take into account the imbalancedness of the classes. Indeed, if we assume the known classes to be representative of the environment, we would expect the proportions of each classes to be preserved in the unknown samples. In others words, we should expect to see $4 - 5$ Callianassa for $19 - 18$ Upogebia. In that case, a simple basic classifier that would always predict the second class would not make more than $4 - 5$ mistakes in this dataset. While this seems reasonable, such a result is completely useless.

The number of features selected as well as the size of the active feature set is found with a cross validation process as described above. As we have a total of 253 known samples, we split them in 23 subsets of size 11 and use a leave-one-set-out cross-validation. For every training set of 230 samples, we selected 170 OTUs as relevant and compare the consistency of these 23 subsets. The average consistency of 0.9952 justifies that we select the final relevant OTUs as the intersection of all feature subsets. Indeed by taking the intersection of subsets we get rid of the OTUs that are highly dependent on the current training set while keeping the ones that are common to all training samples. This yields a final dimensionality of 167 OTUs out of the 7785 to which we have access (this corresponds to a reduction of almost 99% of the dimensionality). Based on this final set of OTUs, we train an NBC and apply it to our unknown samples. It turns out that 7 samples are classified as Upogebia (sample IDs 773.5386, 769.5386, 780.5386, 784.5386, 704.5387, 791.5387 and 807.5386) and the remaining ones are classified as Callianassa. Moreover, the classification results are the same whether we use JMI or MIM.

5 Conclusion

In this work we analyzed the EMP from a data science perspective. Motivated by the problem of predicting missing entries in environmental parameters of metagenomic studies, we investigated the use of greedy feature selection methods combined with information-theoretic criteria. As a consequence we were able to extract relevant OTUs in a human gut that are the most related to the sex of their host. Secondly we derived methods to reduce the number of OTUs for data prediction purposes. Combining our feature selection with a naïve Bayes classifier enables us to classify shrimps from a sample as Callianassa or Upogebia. We provided empirical evidence that using only 1% of the OTUs allowed for accurate prediction.

References


