Using Machine Learning to Identify Major Shifts in Human Gut Microbiome Protein Family Abundance in Disease

Mehrdad Yazdani*†, Bryn C. Taylor‡, Justine W. Debelius‡, Weizhong Li§, Rob Knight¶ and Larry Smarr*‖
*California Institute for Telecommunications and Information Technology, UC San Diego, California, USA
†Open Medicine Institute, Mountain View, California, USA
‡Biomedical Sciences, UC San Diego, California, USA
§J. Craig Venter Institute, La Jolla, California, USA
¶Department of Pediatrics, UC San Diego, California, USA
‖Harry E. Gruber Professor, Department of Computer Science and Engineering, UC San Diego, California, USA
Email: myazdani@ucsd.edu

***Abstract*—Inflammatory Bowel Disease (IBD) is an autoimmune condition that is observed to be associated with major alterations in the gut microbiome taxonomic composition. Here we classify major changes in microbiome protein family abundances between healthy subjects and IBD patients. We use machine learning to analyze results obtained previously from computing relative abundance of ∼10,000 KEGG orthologous protein families in the gut microbiome of a set of healthy individuals and IBD patients. We develop a machine learning pipeline, involving the Kolomogorv-Smirnov test, to identify the 100 most statistically significant entries in the KEGG database. Then we use these 100 as a training set for a Random Forest classifier to determine ∼5% the KEGGs which are best at separating disease and healthy states. Lastly, we developed a Natural Language Processing classifier of the KEGG description files to predict KEGG relative over- or under- abundance. As we expand our analysis from 10,000 KEGG protein families to one million proteins identified in the gut microbiome, scalable methods for quickly identifying such anomalies between health and disease states will be increasingly valuable for biological interpretation of sequence data.**

***Keywords*-microbiome; ibd; kegg; random forest; pca; machine learning;**

I. Introduction

The exponential decline in the cost of next-generation sequencing technology and innovations in bioinformatics approaches have enabled discovery of the detailed microbial ecology of the human body that were heretofore largely unexplored. In whole genome sequencing metagenomics, the genomic DNA present in a sample is first randomly sheared and then sequenced. The output of the Illumina sequencer are "reads" which have ∼100 contiguous DNA bases per read. Here we use samples that have been deeply sequenced (e.g., 100-200 million reads per sample).

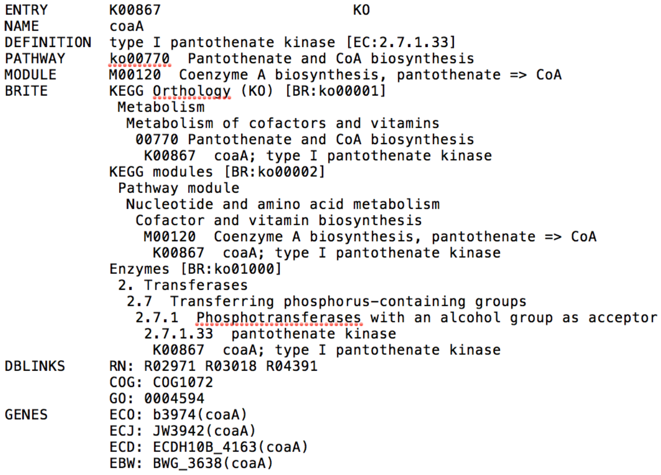
There are 80 autoimmune diseases recognized by the National Institute of Health [1], one of which is Inflammatory Bowel Disease (IBD), which is closely tied to dysbiosis of the gut microbiome [2]. Here we examine deep metagenomic sequencing of a set of healthy subjects and IBD patients to determine how microbial function changes in the health and disease state.

Despite the extensive research on the compositional changes of microbiota in IBD, the precise manner in which changes in the microbial community contributes to the disease state is only beginning to be unraveled. The microbiome DNA contains about 100 times as many genes as its human host DNA, carrying out important functions for the host, such as modulating immune development, amino acid biosynthesis, and energy harvest from food [3].

In previous work [4], [5], the bacterial species compositions were shown to be highly variable across healthy subjects, but the relative abundance of different metabolic pathways were extremely consistent between individuals and over time (see Figure 3 in [5] and Figure 2 in [4]). However, one study [5] included otherwise healthy obese individuals, and the other [4] included a cohort of individuals rigorously defined as healthy at every body site. Testing whether this pattern of constancy of metagenome-encoded functional gene frequency holds true for more acutely diseased populations is therefore of considerable interest.

This type of biological information is contained in the Kyoto Encyclopedia of Genes and Genomes (KEGG), which is used to elucidate microbial function. KEGG is a collection of databases that contain information about genomes, biological pathways, drugs, chemicals, diseases, and protein family functions [6], [7]. In the KEGG database, each entry has a specific K number and describes an orthologous protein family with a particular biological function. Each KEGG also has a text entry such as the one shown in Figure 1 (accessible through the BioServices Python package [8]). Understanding the functional profiles of IBD microbiomes, including their differences in health and disease states, instead of just the taxonomic structure of microbial communities, will help inform drug development and other treatment options for patients.

Several studies have explored the function of the IBD microbiome using the KEGG database. For instance, Morgan



```
ENTRY       K00867                      KO
NAME        coaA
DEFINITION  type I pantothenate kinase [EC:2.7.1.33]
PATHWAY     ko00770  Pantothenate and CoA biosynthesis
MODULE      M00120  Coenzyme A biosynthesis, pantothenate => CoA
BRITE       KEGG Orthology (KO) [BR:ko00001]
             Metabolism
              Metabolism of cofactors and vitamins
               00770 Pantothenate and CoA biosynthesis
                K00867  coaA; type I pantothenate kinase
            KEGG modules [BR:ko00002]
             Pathway module
              Nucleotide and amino acid metabolism
               Cofactor and vitamin biosynthesis
                M00120  Coenzyme A biosynthesis, pantothenate => CoA
                 K00867  coaA; type I pantothenate kinase
            Enzymes [BR:ko01000]
             2. Transferases
              2.7  Transferring phosphorus-containing groups
               2.7.1  Phosphotransferases with an alcohol group as acceptor
                2.7.1.33  pantothenate kinase
                 K00867  coaA; type I pantothenate kinase
DBLINKS     RN: R02971 R03018 R04391
            COG: COG1072
            GO: 0004594
GENES       ECO: b3974(coaA)
            ECJ: JW3942(coaA)
            ECD: ECDH10B_4163(coaA)
            EBW: BWG_3638(coaA)
```

Figure 1: An example of a KEGG description file as queried from the KEGG database for K00867.

et al. [9] created a broad map with 16S sequences of the gut microbiota of a large cohort of patients with IBD, and then chose a representative 11 samples to perform metagenomic sequencing and analyze with the KEGG database. This analysis revealed that moderate perturbation of microbiome composition corresponds with major perturbation of metabolic and functional pathways. Greenblum et al. [10] developed a metabolic network of KEGG enzymes to study the enzymatic variation in the gut microbiome of patients with IBD. Tong et al. [11] identified functional microbial communities using 16S rRNA sequencing, enhancing the analysis with reference sequences from Greengenes and then annotating the predicted genes with the KEGG database. Erickson et al. [12] found a number of KEGGs and KEGG pathways that were altered in Ileal Crohn's Disease.

In our previous work [13], we used deep metagenomic sequencing data, instead of predicting genes from 16S sequencing data, to compute relative abundances of the entire ∼10,000 entry KEGG database. Here we extend these results, using machine-learning techniques to discover the most significant over- and under- abundant KEGGs in the disease state compared to healthy subjects.

In section II (Previous Work), we discuss our data collection process and previous results. In section III (Methods), we present our proposed algorithms and workflows. We have two specific workflows. The first workflow is to identify KEGGs that are over or under abundant in disease states based on relative abundance data obtained from stool samples from healthy and disease cohorts. In the second workflow we train an NLP classifier using the KEGG description files to predict over and under abundant set of KEGGs that we identified from our first workflow. In section IV (Results) we show the results and evaluate our proposed workflows and we conduced in section (V).

II. Previous work

A. Cohort selection and data extraction

The three main subtypes of IBD are Ileal Crohn's Disease (ICD), Colonic Crohns Disease (CCD), and Ulcerative Colitis (UC) [14]. In our earlier research we developed a study with examples of each subtype of IBD, as well as a set of healthy subjects. The description of the set of individuals is contained in our earlier paper [13]. In summary, we downloaded 2.4 TBs of raw reads from 34 healthy individuals, 6 samples from UC and 15 from ICD selected from the NIH National Center for Biotechnology Information (NCBI) BioProjects 46321, 46881 and 43021. An additional seven samples were deeply sequenced (200 million reads per sample) by the J. Craig Venter Institute from an adult with early CCD. Table I summarizes our dataset and cohort nomenclature.

Table I: Cohort sample distribution

Sample distribution for the various cohorts in our dataset.

Cohort	Abbreviation	Number of Samples
Healthy subjects	HE	34
Ulcerative colitis	UC	6
Ileal Crohn's disease	CD	15
Colonic Crohn's disease	LS	7
	Total samples:	62

B. Feature annotation

To clarify how our previous study computed the KEGG relative abundances across our patient set, we describe the technical process we followed. First, we created a reference database of known (as of Sept 2012) gut microbe genomes consisting of 2,471 complete and 5,543 draft Bacterial and Archaeal genomes, 2,399 complete virus genomes, 26 complete Fungal genomes, and 309 HMP Eukaryote Reference Genomes, for a grand total of 10,012 genomes representing 30GB of sequences. We then used the San Diego Supercomputer Center's Gordon supercomputer to align our 6.4 billion reads from the healthy and IBD samples against the reference database we created. The database was used to calculate the relative taxonomic distribution in each sample.

In addition, high quality filtered reads were also assembled into contigs (using Velvet [15]) and Open Reading Frames (ORFs) were predicted from the contigs using Metagene [16]. Protein families were identified using the KEGG database [6], [7]. All the ORFs were aligned to KEGG sequence database using BLASTP. A curated KEGG reference database was generated by clustering all KEGG sequences at 90% sequence identity with CD-HIT [17]. If all sequences in a CD-HIT cluster belonged to the same protein orthology family (KO), the longest representative sequence was used in the reference database; otherwise all sequences were retained.

The curated database recovered more than 99% of the original hits and was 10 times faster [18]. Only the top score non-overlapping alignments from the ORFs to KEGG BLAST alignment results were used in counting the KEGG protein abundance. KEGG abundance was calculated as the number of times a KEGG protein is found in a sample, normalized against the reference protein length and predicted ORF length. The abundance of a protein family was calculated as its abundance divided by the sum of the abundance of all protein families.

The computations described above consumed 180,000 core-hours (provided by Director Michael Norman) on the Gordon supercomputer at the San Diego Supercomputer Center. About half of this time was required for the KEGG analysis.

The resulting output was a database of 10,012 KEGG entries with relative abundance for each KEGG for each of the 62 human gut microbiome samples in Table I. This database was completed in August 2014 and is available upon request.

III. Methods

The dataset we examine represents a matrix of $10,012 \times 62$ or 620,744 entries. This is the matrix on which we use machine learning techniques to ascertain if there are biomedically relevant patterns in this dataset. In the near future not only will the number of samples increase by an order of magnitude, but we will be able to compute directly the relative abundance of $\sim$1 million genes, or two orders of magnitude over our current KEGG dataset. Thus, within a year, we expect our matrix to grow by three orders of magnitude in scale. This paper is our pilot to develop machine learning algorithms which will scale with the increase in data size.

First, we use Principal Component Analysis (PCA) to investigate our data set both across samples and across KEGGs. That is, we apply PCA to both the $62 \times 10,012$ matrix (a PCA across samples) and to the transpose matrix (thus a $10,012 \times 62$ matrix, or a PCA across KEGGs). This reveals insights into the structure of the data from both a samples and KEGGs perspective.

Second, we develop a KEGG relative abundance classifier that predicts over or under abundant KEGGs in the disease state compared to the healthy. Third, having classified all KEGGs, we then train a classifier using the queries from the KEGG database to predict if a KEGG is over or under abundant in disease state only using the text in the description file (example description file is shown in Figure 1). Note that we deploy this 2-step process since currently ground truth for which KEGGs are over or under abundant in disease state is not well understood.

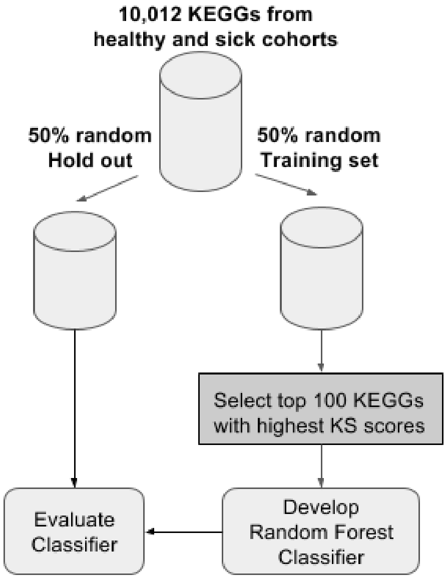

10,012 KEGGs from healthy and sick cohorts
50% random Hold out
50% random Training set
Select top 100 KEGGs with highest KS scores
Evaluate Classifier
Develop Random Forest Classifier

Figure 2: Workflow for developing a Random Forest classifier to discriminate between over and under abundant KEGGs in the diseased state. After splitting the data into the training set, we select the top 100 KEGGs that have the highest Kolomogorv-Smirnov (KS) score. We then use these 100 most significant KEGGs to train a Random Forest classifier and use to predict and evaluate on the remainder of the dataset.

A. Discrimination between healthy and IBD cohort using relative abundances

The workflow for discriminating between healthy and IBD cohorts in our samples uses the Kolmogorov-Smirnov (KS) test and Random Forests. We chose the KS test since it does not use any assumptions on the distribution of the data. We chose Random Forests since such classifiers are scale invariant, non-linear, and robust to outliers, missing values, and overfitting. Note that the overall design of the workflow is our main aim, not the specific choice of classifier (Random Forest) and statistical test (KS).

As indicated in Figure 2, we first randomly partition the $\sim$10,000 KEGGs into a 50% hold-out set and a 50% training set. We develop our classifier from the training set and apply the classifier on all KEGGs (the union of the training and hold-out sets). The primary reason for splitting the KEGGs into two sets is to simulate the scenario of scaling our workflow to new KEGGs that are introduced in the database. In other words, we simulate, in a controlled manner, the issues with developing machine learning models from databases that grow larger over time. Thus, our workflow assumes that we are training our model with a database that is smaller than the time of the application of the model.

Since there is no single ground truth on which KEGGs should be over and under abundant in disease state (as is common in many biological datasets), from the training set we use the KS test to determine the subset of the KEGGs that are the most statistically significant between the disease and healthy cohorts. From the KS test we select the 100 KEGGs with the highest KS scores, and determine whether these are over or under abundant in IBD cohorts. The over or under abundance is determined by comparing the median of the abundance for each of the 100 KEGGs between the healthy and IBD cohorts. In short, we have determined the most statistically significant over and under abundant KEGGs in disease state.

Using these 100 most significant KEGGs, we train a Random Forest classifier (defaults of RF, using 500 trees, sampling with replacement, and the square root of the number of predictors) to model if a KEGG is over or under abundant relative to the healthy state. With this classifier we can also compute a confidence score (probability that a KEGG is under abundant in the disease cohorts) as estimated by the Random Forest model to all the KEGGs in our data. This is possible as we report the KEGGs that have the highest (corresponding to under abundance of IBD cohorts relative to healthy) and lowest (corresponding to over abundance of IBD cohorts relative to health) confidence scores. Note that we take this approach of only training the Random Forest classifier on 100 KEGGs and applying to the remaining KEGGs to ensure that we do not overfit.

B. Discrimination between healthy and IBD cohorts using KEGG description files

Once we have classified all KEGGs as either over or under abundant, we can now develop a classifier that determines if a KEGG is over or under abundant in a subject based on the KEGG description file (example of a description file for one KEGG is shown in Figure 1). Since the KEGG description file is a text file and numerous approaches and methods are available for Natural Language Processing, we here present a baseline model to serve as a benchmark for future models. In our baseline approach, we use the "raw" KEGG description file as queried from the database [6], [7]. In this baseline approach, we extract bag-of-words unigram features weighted by Term Frequency-Inverse Document Frequency scaling (TF-IDF, as implemented in [19]), thus ignoring the word order and hierarchical structure of the description file (more sophisticated features, such as [20], [21], can be explored for future work). Our workflow for this approach is shown in Figure 3.

In the bag-of-words approach to NLP, in each document we count the number of occurrences of the terms in our dictionary (fixed-size vocabulary), creating a count vector that we can use as a feature vector for machine learning tasks. The idea here is that high occurrences of specific terms reflect the content or subject of the document. These raw counts of terms are referred to as "term frequencies". Since the raw counts of the terms may inadvertently weight so-called "stop words" (such as "the") that do not reflect the subject of a document, we normalize the raw counts of each term to diminish bias from commonly occurring words. A common normalization is to count the number of documents that contain each of the terms in our dictionary, and this count is referred to as "document frequency." The motivation with this normalization is that if a word is common then it should appear in most of the documents in the corpus that we are studying and will have a high document frequency count. We use the document frequencies to normalize the term frequencies to obtain the TF-IDF features.

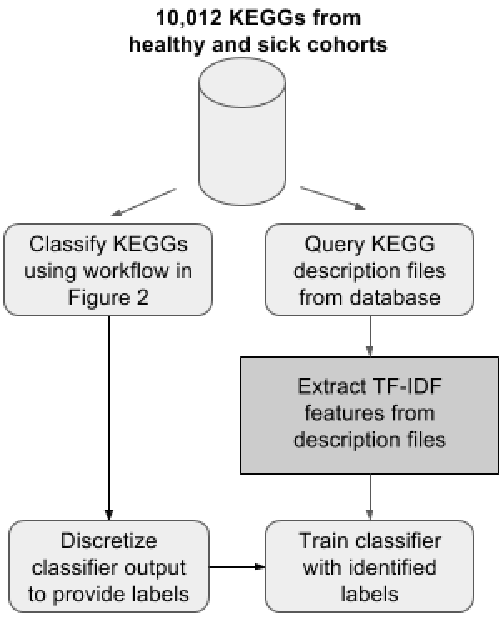

10,012 KEGGs from healthy and sick cohorts
Classify KEGGs using workflow in Figure 2
Query KEGG description files from database
Extract TF-IDF features from description files
Discretize classifier output to provide labels
Train classifier with identified labels

Figure 3: Workflow for developing an NLP classifier to discriminate between over and under abundant KEGGs in the diseased state using the classified labels as determined by the workflow from Figure 2. Each of the 10,012 KEGGs are queried in the KEGG database [7], [6] and stored as KEGG description files. We then extract Term Frequency-Inverse Document Frequency (TF-IDF) features from the description files as described in Equations 1 and 2. In parallel we classify each of the 10,012 KEGGs from our patient data using the workflow in Figure 2 and discretize the probabilities of the classifier into over and under abundant and "neutral" categories. These categories are then used as labels to train a classifier on the extracted TF-IDF features from the KEGG description file.

In our application, we define each of the terms (also referred to as word or token) that occur in the KEGG description files as t and each of the KEGG description files as d (also referred to as a document in NLP). Our corpus of documents is the 10,012 KEGG description files and we compute the TF-IDF features for each KEGG as

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \cdot \big(\text{idf}(t,d) + 1\big) \tag{1}$$

where $\text{tf}(t,d)$ is the number of occurrences (that is, the frequency) of term t in the KEGG description file d and $\text{idf}(t,d)$ is the normalization of this count with respect to the number of occurrences of t in all the other KEGG description files. This normalization $\text{idf}(t,d)$ is computed as:

$$\text{idf}(t,d) = \log \frac{1+N}{1+\text{df}(d,t)} \tag{2}$$

where N is the total number of KEGG description files (10,012), and $\text{df}(d,t)$ is the number of KEGG description files with the term t. Given the success of such features in NLP applications, this set of features from the description files for each KEGG serves as a baseline to develop classification models.

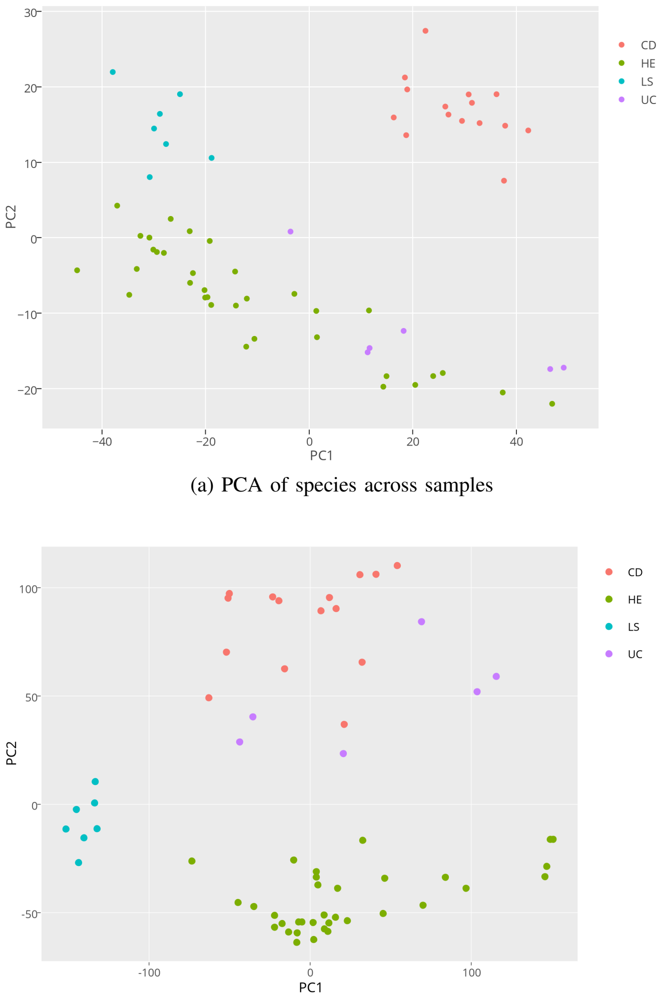
To develop the classification models from the bag-of-words TF-IDF features, we discretize the confidence scores of all the KEGGs in our data, as estimated by the Random Forest classifier we developed earlier, to three distinct categories: under abundant, over abundant, and neither (probability of Random Forest output greater than 0.75, less than 0.25, and between 0.25 and 0.75 respectively. These thresholds may be adjusted to allow for more False Positives/Negatives, but we do not explore adjusting these trade-offs here. Using these categories as the labels for the bag-of-words features that we computed, we proceed to train three standard classifiers (Naive Bayes, Support Vector Machine with linear kernel, and Logistic Regression) and report their average F1 score using 10-fold Cross Validation.

IV. Results

A. Use of PCA to show KEGGs separate healthy and disease states

It has been known since 2010 (see Figure 4 in [22]) that in human gut microbiome samples the species abundances can separate healthy from IBD substates UC and ICD using PCA. Further research (see Figure 2 in [23]) showed that PCAs based on species abundance can also separate ICD into its two subtypes (CCD and ICD), as well as separating UC and healthy.

Our species abundance data shows a similar separation (Figure 4a). However, we can go beyond microbial species and use PCA on the KEGG protein families by computing PCA on our data matrix with microbiome samples as rows and KEGGs as columns. This shows an even clearer separation between healthy and the three disease states (Figure 4b). Thus, it appears that there are significant differences between the KEGG relative abundances in health and each of the three IBD disease states. We next turn to using machine learning techniques to find which KEGGs are the best discriminators.



(a) PCA of species across samples

(b) PCA of KEGGs across samples

Figure 4: PCA of species (a) and KEGGs (b) across samples colored by the different cohorts (abbreviations and data summary shown in Table I). As shown in (b), using all 10,012 KEGGs we see near perfect separation between the different cohorts. While on the other hand, in (a) using species PCA we do not see a clear separation between UC (Ulcerative colitis) and HE (healthy) groups.

B. Classification of over and under abundant KEGGS in IBD

The Random Forest classifier that we developed according to our workflow in Figure 2 obtains an out-of-bag classification accuracy of 99%. Moreover, since Random Forest classifiers can give probabilistic outputs, we compute the confidence scores for how well each KEGG works as a classifier on separation of over or under abundant compared to the healthy cohort. Figure 5 shows the PCA of the KEGGs. We color each KEGG in the PCA scatter plot with the discretized confidence score from the Random Forest classifier[1].

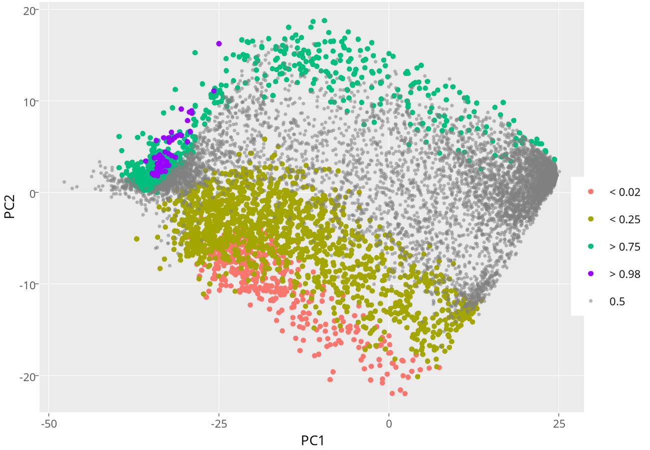


Figure 5: PCA of KEGGs classified with the outputs of our trained Random Forest classifier based on subject relative abundance data. The categories of over abundant, under abundant and neutral KEGGs form coherent clusters in PCA space suggesting that the classifier has not overfit and that similar KEGG distributions in our patient population are classified with a similar label.

Coloring the PCA plot with these discretized confidence scores shows that the all KEGGs are clustered according to their abundance level. That is, KEGGs that are clustered with each other in PCA space have a similar classification score as given by the Random Forest classifier. Otherwise if the classifier was overfitting, then we would expect to see the color distribution of the KEGGs in the scatter plot to be distributed in a less structured and more random manner. Note also that this separation continues into the most extreme regions of confidence (over 98% and less than 2%). This set of "extreme" KEGGs corresponds to $\sim$500 KEGGs or 5% of all KEGGs in our dataset. These results indicate that our proposed workflow in Figure 2 is able to discover the best KEGG classifiers for separating over or under abundant values in IBD patients compared to healthy subjects.

Finally, we can use the classifier to find the over- and under-abundance KEGGs that most differentiate between the health and disease states. In Figure 6 we show a sample of the most confident KEGGs that are over and under abundant compared to the healthy cohorts for both our training and hold-out sets. This figure clearly shows the separation of the healthy and disease samples on a logarithmic scale.

It is beyond the scope of this paper to go into the biological implications of these large differences, but we note that our machine learning methodology has selected certain KEGGs that previous research has identified as important to IBD state. Notably, we find that a number of the over-abundant KEGGs identified are involved in the phospho-transferase system (PTS) (K03480, K03483, K03475, K02794, and others). The PTS is a sugar transport mechanism associated with the Firmicutes phylum, which is favored in patients with IBD [24], [10], being involved in carbohydrate uptake. Furthermore, the PTS enzyme FrvX is a known biomarker for IBD according to [25].

Another interesting over-abundant KEGG we identify is mobB (K03753), the presence of which enhances activation of nitrate reductase as discussed in Palmer et al. and Eaves et al. [26], [27]. Nitrate reduction is a critical process that produces nitric oxide, which is not synthesized by the human genome. Increased levels of nitric oxide is associated with inflammation, cancer, and IBD as several studies have shown [10], [28], [29], [30]. Under-abundant KEGGs include those that metabolize amino acids and carbohydrates (K01847, K01711, K00971, K12111, and more), which are thought to be decreased in favor of nutrient uptake in the IBD microbiome as shown in [9]. Several KEGGs involved in amino acid biosynthesis and carbohydrate metabolism are also over-abundant, reflecting the inconsistency in previous studies and the need for further analysis and more datasets.

In a future paper, we will analyze in more detail the biological significance of our hundreds of over and under abundant KEGGs that differentiate between health and IBD.

C. Development of a natural language classifier to disease association

As additional KEGGs are annotated or additional disease pathways are identified, natural language processing can help predict the association between new protein families and disease. Here, we present preliminary results on developing a baseline classifier that determines if a KEGG is over or under abundant based on the KEGG description file alone (a snippet is shown in Figure 1).

As discussed in the methods section above, we extract unigram bag-of-words TF-IDF features and train three different baseline classifiers to classify a KEGG as "over abundant," "under abundant," or "neither" based on the results of the two-stage classification above. Figure 5 shows the distribution of these three categories on the PCA of the KEGG relative abundances amongst the cohorts in our data. The uniformity of the distribution of these three categories in the distribution of our KEGGs suggests that the categories that we have selected for the KEGGs are sensible.

Table II shows the average F1 score for the three classifiers we considered using 10-fold cross validation. We report the F1 score since it is a more rigorous measurement of accuracy as it is the harmonic mean of the precision and recall scores (that is, we are accounting for both type I and type II errors by not allowing class imbalance to influence our error rates). The results that we have are significantly

[1] For the raw scores see https://plot.ly/∼crude2refined/1959/pc2-vs-pc1.embed

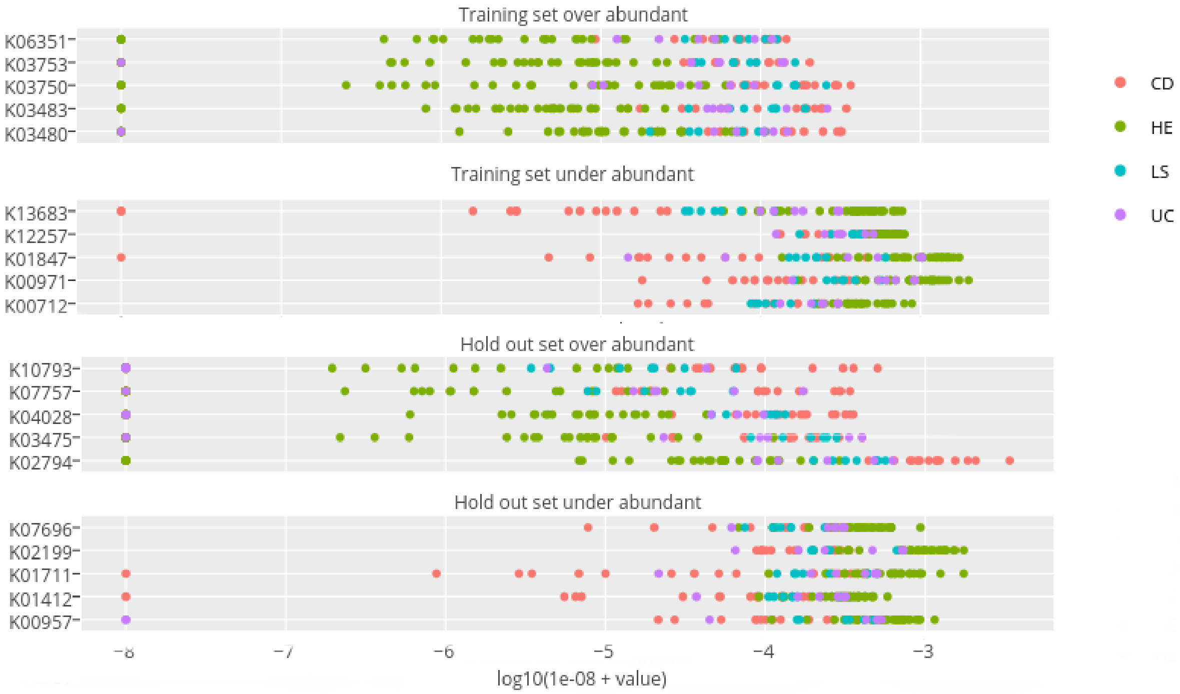


Figure 6: Distribution of the relative abundance of KEGGs selected by our approach that discriminate between healthy and disease states. The horizontal axis is the relative abundance values of the KEGGs on a logarithmic scale for each of the samples. See Table I for the summary of cohort samples.

Table II: Classification accuracies using KEGG description files

Average F1 scores based on 10-fold cross validation for classifiers we trained on bag-of-words TF-IDF features from the KEGG description files to classify KEGGs as under or over abundant relative to healthy states.

Classifier	F1-score
Naive Bayes	0.71 ± 0.024
Support Vector Machine	0.76 ± 0.012
Logistic Regression	0.77 ± 0.007

higher than random. This suggests that our baseline TF-IDF features are able to predict if the description of the KEGGs has predictive power for discriminating between healthy and disease states. The predictive power of these features can be used in more sophisticated topic modelling approaches (such as Non-Negative Matrix Factorization [31], Latent Dirichlet Allocation [32], and word embeddings [20], [21]). Such topic models can then be used to aid biologists in comprehending the biological relationship between the numerous KEGGs in databases. We suspect that future work that takes into account the hierarchical structure and utilizes domain knowledge of the KEGG description files will significantly improve this baseline performance.

V. Discussion and Related Work

Microbial communities are complex networks that rely on function as well as structure. The KEGG is a well-characterized database of molecular function that has been a widely-used tool to investigate microbial function. By looking at the function of specific disease-associated microbial communities, we can better identify targets for future intervention (i.e. small molecule development to target a specific gene pathway). The motivation for using machine learning methods is to reduce the amount of time-consuming manual investigation of immense amounts of data generated from metagenomic sequencing. Using metagenomic data from a cohort of healthy and IBD-affected individuals, we developed and trained a two step classifier to identify 100 KEGG ortholog genes which are over or under abundant in IBD patients compared to healthy adults. We also demonstrated the ability of a simple natural language classifier to identify KEGGs as over or under abundant in the IBD disease state.

While there are a number of methods that could be used with KEGG protein families to discover the large changes in healthy and disease states, we turned to machine learning methods here because the next step in our project will see our matrix of samples versus function grow by a factor of 1000x, necessitating a computational approach to discovery of these patterns.

Acknowledgments

The authors would like to thank members of the UC San Diego Knight Lab and Calit2 Qualcomm Institute staff for many useful conversations. We also thank the San Diego Supercomputer Center for providing computational resources. We also acknowledge support from the Amar Foundation.

References

[1] "NIH Autoimmune Diseases. Bethesda, Maryland." www.niaid.nih.gov/topics/autoimmune, accessed: 2016-05-15.

[2] W. A. Walters, Z. Xu, and R. Knight, "Meta-analyses of human gut microbes associated with obesity and ibd," *FEBS letters*, vol. 588, no. 22, pp. 4223–4233, 2014.

[3] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project: exploring the microbial part of ourselves in a changing world," *Nature*, vol. 449, no. 7164, p. 804, 2007.

[4] H. M. P. Consortium *et al.*, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[5] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit *et al.*, "A core gut microbiome in obese and lean twins," *nature*, vol. 457, no. 7228, pp. 480–484, 2009.

[6] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Kegg as a reference resource for gene and protein annotation," *Nucleic acids research*, vol. 44, no. D1, pp. D457–D462, 2016.

[7] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[8] T. Cokelaer, D. Pultz, L. M. Harder, J. Serra-Musach, and J. Saez-Rodriguez, "Bioservices: a common python package to access biological web services programmatically," *Bioinformatics*, vol. 29, no. 24, pp. 3241–3242, 2013.

[9] X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper *et al.*, "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment," *Genome Biol*, vol. 13, no. 9, p. R79, 2012.

[10] S. Greenblum, P. J. Turnbaugh, and E. Borenstein, "Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease," *Proceedings of the National Academy of Sciences*, vol. 109, no. 2, pp. 594–599, 2012.

[11] M. Tong, X. Li, L. W. Parfrey, B. Roth, A. Ippoliti, B. Wei, J. Borneman, D. P. McGovern, D. N. Frank, E. Li *et al.*, "A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease," *PloS one*, vol. 8, no. 11, p. e80702, 2013.

[12] A. R. Erickson, B. L. Cantarel, R. Lamendella, Y. Darzi, E. F. Mongodin, C. Pan, M. Shah, J. Halfvarson, C. Tysk, B. Henrissat *et al.*, "Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of crohn's disease," *PloS one*, vol. 7, no. 11, p. e49138, 2012.

[13] S. Wu, W. Li, L. Smarr, K. Nelson, S. Yooseph, and M. Torralba, "Large memory high performance computing enables comparison across human gut microbiome of patients with autoimmune diseases and healthy subjects," in *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*. ACM, 2013, p. 25.

[14] I. Cleynen, G. Boucher, L. Jostins, L. P. Schumm, S. Zeissig, T. Ahmad, V. Andersen, J. M. Andrews, V. Annese, S. Brand *et al.*, "Inherited determinants of crohn's disease and ulcerative colitis phenotypes: a genetic association study," *The Lancet*, vol. 387, no. 10014, pp. 156–167, 2016.

[15] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de bruijn graphs," *Genome research*, vol. 18, no. 5, pp. 821–829, 2008.

[16] H. Noguchi, T. Taniguchi, and T. Itoh, "Metageneannotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes," *DNA research*, vol. 15, no. 6, pp. 387–396, 2008.

[17] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "Cd-hit: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[18] S. Wu, Z. Zhu, L. Fu, B. Niu, and W. Li, "Webmga: a customizable web server for fast metagenomic sequence analysis," *BMC genomics*, vol. 12, no. 1, p. 1, 2011.

[19] 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*, vol. 12, pp. 2825–2830, 2011.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[21] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–43.

[22] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing," *nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[23] B. P. Willing, J. Dicksved, J. Halfvarson, A. F. Andersson, M. Lucio, Z. Zheng, G. Järnerot, C. Tysk, J. K. Jansson, and L. Engstrand, "A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes," *Gastroenterology*, vol. 139, no. 6, pp. 1844–1854, 2010.

[24] M. A. Mahowald, F. E. Rey, H. Seedorf, P. J. Turnbaugh, R. S. Fulton, A. Wollam, N. Shah, C. Wang, V. Magrini, R. K. Wilson *et al.*, "Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla," *Proceedings of the National Academy of Sciences*, vol. 106, no. 14, pp. 5859–5864, 2009.

[25] C.-S. Chen, S. Sullivan, T. Anderson, A. C. Tan, P. J. Alex, S. R. Brant, C. Cuffari, T. M. Bayless, M. V. Talor, C. L. Burek *et al.*, "Identification of novel serological biomarkers for inflammatory bowel disease using escherichia coli proteome chip," *Molecular & Cellular Proteomics*, vol. 8, no. 8, pp. 1765–1776, 2009.

[26] T. Palmer, C.-L. Santini, C. Iobbi-Nivol, D. J. Eaves, D. H. Boxer, and G. Giordano, "Involvement of the narj and mob gene products in distinct steps in the biosynthesis of the molybdoenzyme nitrate reductase in escherichia coli," *Molecular microbiology*, vol. 20, no. 4, pp. 875–884, 1996.

[27] D. J. Eaves, T. Palmer, and D. H. Boxer, "The product of the molybdenum cofactor gene mobb of escherichia coli is a gtp-binding protein," *European Journal of Biochemistry*, vol. 246, no. 3, pp. 690–697, 1997.

[28] G. Kolios, V. Valatas, and S. G. Ward, "Nitric oxide in inflammatory bowel disease: a universal messenger in an unsolved puzzle," *Immunology*, vol. 113, no. 4, pp. 427–437, 2004.

[29] G.-Y. Yang, S. Taboada, and J. Liao, "Induced nitric oxide synthase as a major player in the oncogenic transformation of inflamed tissue," *Inflammation and Cancer: Methods and Protocols: Volume 2: Molecular Analysis and Pathways*, pp. 119–156, 2009.

[30] S. E. Winter, C. A. Lopez, and A. J. Bäumler, "The dynamics of gut-associated microbial communities during inflammation," *EMBO reports*, vol. 14, no. 4, pp. 319–327, 2013.

[31] M. W. Berry and M. Browne, "Email surveillance using non-negative matrix factorization," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 249–264, 2005.

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.