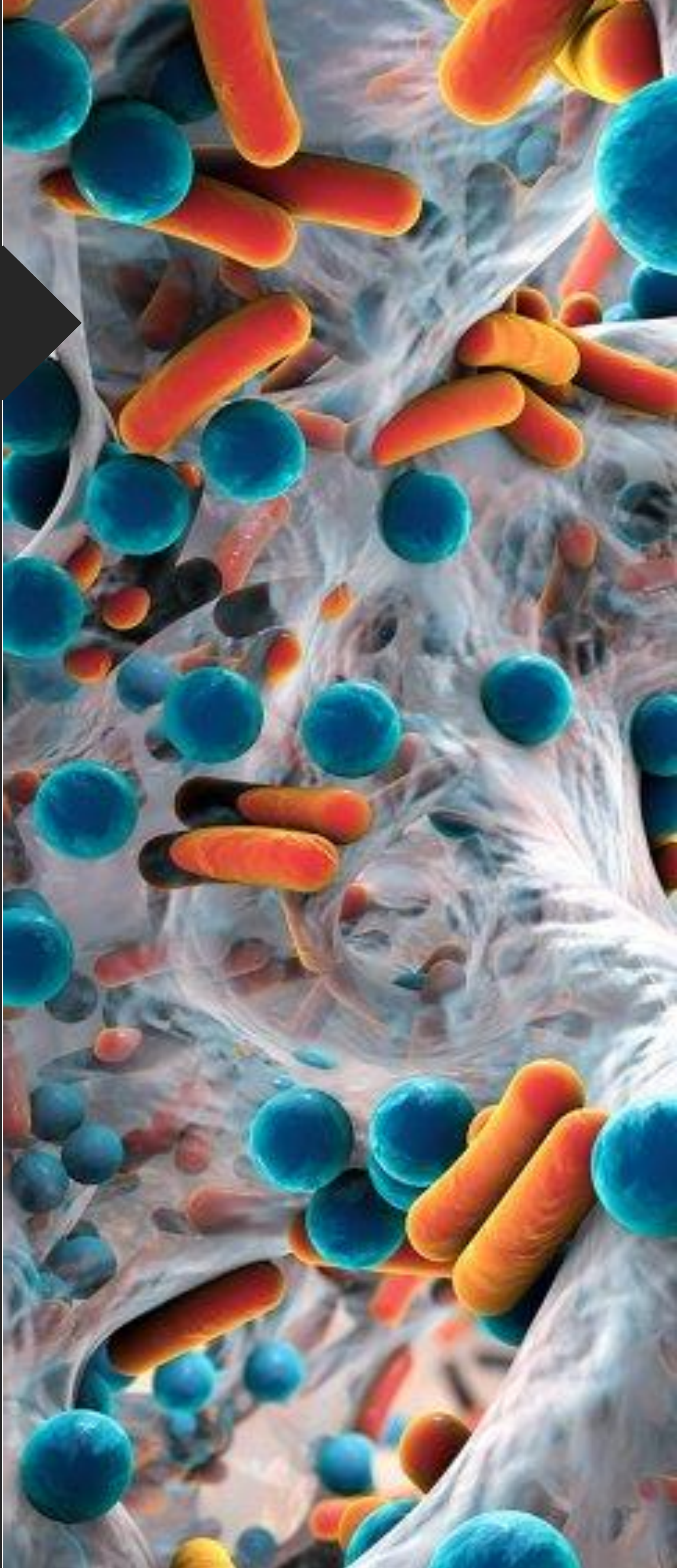
PATTERN
COMPUTER®

EXTRACTING INSIGHTS ON THE DYNAMIC HEALTH-DISEASE TRANSITIONS IN THE HUMAN GUT MICROBIOME



Larry Smarr,[1-3] Marc Jaffrey,[4] Michael Dushkoff,[4] Brynn Taylor,[5] Pilar Ackerman,[4] Mehrdad Yezdani,[3] Weizhong Li[3,6]

[1]Center for Microbiome Innovation, University of California, San Diego, San Diego, California, USA. [2]Department of Computer Science and Engineering, University of California, San Diego, San Diego, California, USA. [3]California Institute for Telecommunications and Information Technology, University of California, San Diego, San Diego, California, USA. [4]Pattern Computer Inc., 38 Yew Lane, Friday Harbor, WA 98250. [5]Department of Biomedical Sciences, University of California San Diego, La Jolla, CA, USA. [6]Center for Research on Biological Systems, University of California San Diego, California, USA

Pattern Computer Inc.

© 2018 Pattern Computer, Inc. All Rights Reserved.

No part of this publication may be reproduced, or transmitted, in any form or by any means, mechanical, electronic, photocopying, recording, or otherwise, without prior written permission of Pattern Computer Inc., unless it is for research or educational purposes in which case no such approval is required.

No licenses, express or implied, are granted with respect to any of the technology described in this document. Pattern Computer Inc. retains all intellectual property rights associated with the technology described in this document. This document is intended to inform about Pattern Computer product offerings and technologies and its implementations.

Pattern Computer Inc.
38 Yew Lane, Friday Harbor, WA 98250.
USA

PATTERN COMPUTER MAKES NO WARRANTY OR REPRESENTATION, EITHER EXPRESS OR IMPLIED, WITH RESPECT TO THIS DOCUMENT, ITS QUALITY, ACCURACY, MERCHANTABILITY, OR FITNESS FOR A PARTICULAR PURPOSE. AS A RESULT, THIS DOCUMENT IS PROVIDED “AS IS,” AND YOU, THE READER, ARE ASSUMING THE ENTIRE RISK AS TO ITS QUALITY AND ACCURACY.

IN NO EVENT WILL PATTERN COMPUTER BE LIABLE FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES RESULTING FROM ANY DEFECT, ERROR OR INACCURACY IN THIS DOCUMENT, even if advised of the possibility of such damages.

Some jurisdictions do not allow the exclusion of implied warranties or liability, in which case the above exclusion do not apply.

Contents

Key Insights

- Some Inflamatory bowel disease (IBD) may be associated with abnormal oxidative phosphorylation or oxidative stress.
- Identified a dynamic path from disease to health states In human gut microbiome.

Keywords

human gut microbiome, Inflamatory bowel disease (IBD), KEGG, machine learning, t-SNE, principal component analysis

“The crucial first step to data analytics is subspace selection. If you get to the right subspace, everything else is likely to be easier.”

Abstract

The trillions of microbes living in our large intestine — the gut microbiome — play a profound role in human health and disease [1]. While much has been done to explore its diversity, a full understanding of how the dynamical evolution of the microbiome ecology influences healthy and disease states is only beginning to be understood [2].

In this article, we start by reviewing previous research results, [3] examining the gut microbiome *taxonomic* differences between healthy people and of those suffering from the three subtypes of the autoimmune Inflammatory Bowel Disease (IBD): Ileal Crohn's, Colonic Crohn's, and Ulcerative Colitis [6]. This study was recently expanded to understand the *functional* differences by using the Kyoto Encyclopedia of Genes and Genomes (KEGG) [4] protein families in the gut microbiome of the samples. Each of the entries in the KEGG database describes an orthologous protein family, which have specific biological functions. In the study relative abundances of 10,192 KEGG protein families were computed from the sequencing of the human stool samples. Using traditional machine learning techniques, it was shown [5] that a subset of the KEGG protein families can distinguish between healthy and the IBD states.

In this paper, we describe the results obtained with the Pattern Computer proprietary algorithms, tools, and techniques, using an approach without prior assumptions on this large dataset of 62 human microbiome samples, each with the relative abundance of the ~10,000 KEGG protein families. We identified 39 KEGG protein families that were significant in differentiating the disease states from each other and from healthy states, with 9 of the KEGG protein families (out of ~10,000 total) being most associated with a dynamic path from disease to health in the human-gut microbiome. With the Pattern Computer approach, we reduced the size of the dataset to be analyzed by three orders of magnitude. The biochemical pathways, that 6 out of the 9 KEGG protein families are associated with, suggest a hypothesis for further study: Inflammatory bowel disease (IBD), like other inflammatory diseases, may be associated with abnormal oxidative phosphorylation or oxidative stress.

Computing the Data Set

Using techniques outlined in [3], we obtained the deep metagenomic sequencing (50-200 million Illumina short reads per sample) of 34 different healthy patients (a subset of the NIH Human Microbiome Program) and of 28 samples from patients with the three classes [6] of IBD: Ileal Crohn's (ICD), Colonic Crohn's (CCD), and Ulcerative Colitis (UC), listed in Table 1. In the CCD case, there are 7-time samples over a year and a half from one individual. In the ICD, there are 5 individuals, each with 3 samples collected at six-month intervals. In UC, there are 2 individuals, one with a single sample and the other patient with 5 samples. For the patient with 5 samples, 3 of them are from luminal aspirate and 2 from mucosal biopsy. The 3 are uneven in time with the first two separated by two weeks and the third 4 weeks after the second sample.

Table 1: Cohort sample distributions.

COHORT	ABBREVIATION	NUMBER OF SAMPLES
Healthy subjects	HE	34
Ulcerative colitis	UC	6
Ileal Crohn’s disease	ICD	15
Colonic Crohn’s disease	CCD	7
	Total samples:	62

The 6.4 billion Illumina short reads from the healthy and IBD samples were converted to relative abundances for the taxonomy and the KEGG protein families, by a software system developed by Weizhong Li (Figure 1 reproduced from [3]), utilizing the San Diego Supercomputer Center’s Gordon supercomputer, consuming around 180,000 core-hours or ~25 CPU-Years [3].

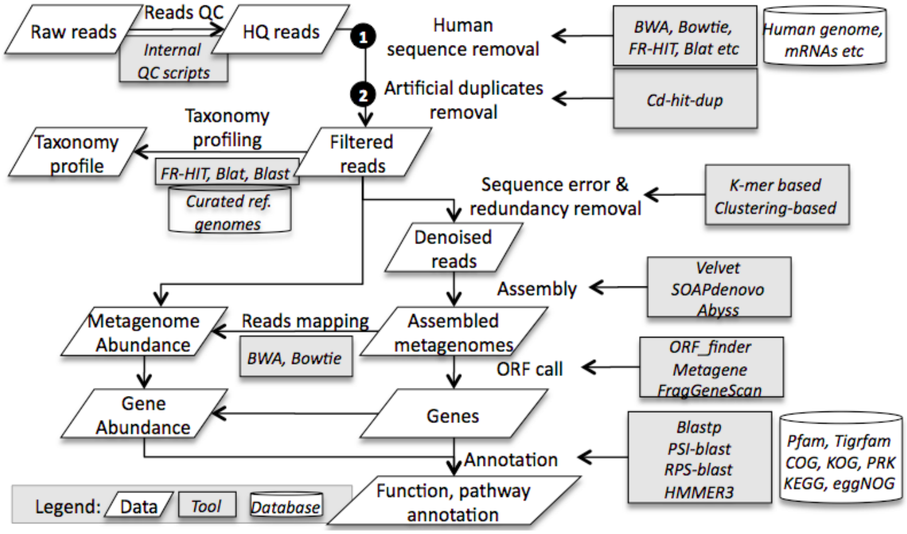


Figure 1: Read-based and assembly-based workflows for Illumina metagenomic data [Figure from reference 3].

Our hypothesis, based on previous published research [7], is that there would be a large difference in the microbiome ecology in these four cases. When we look at the relative abundances of the microbial phyla across the 62 samples (Figure 2), we see that while there is a variation within the disease cohort, the four classes appear to be quite distinct in their ecological composition. The healthy patients are 90+% a mixture of *Bacteroidetes* and *Firmicutes*. In contrast, the ICD are a mixture of *Actinobacteria* and *Firmicutes*, while in UC, *Proteobacteria* is a major admixture with *Bacteroidetes* and *Firmicutes*. The CCD samples seem to be a time varying combination of *Firmicutes, Proteobacteria, Actinobacteria*, and *Euryarchaeota*, with *Bacteroidetes* suppressed in all but one sample.

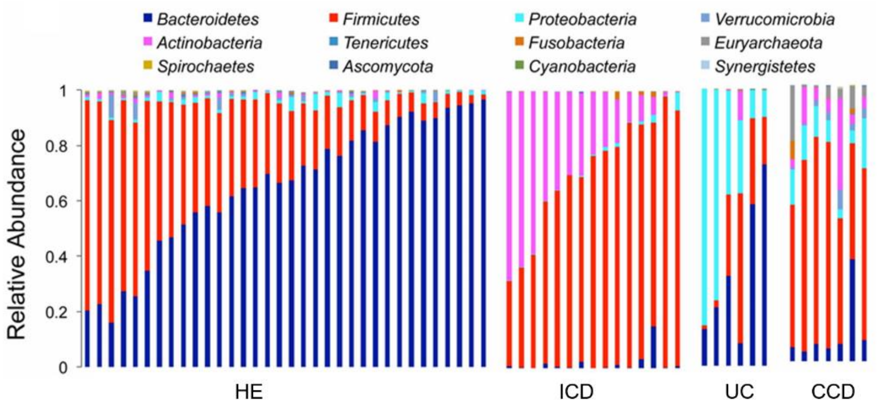


Figure 2: Relative abundance for the samples at phylum level

This observation is verified when we carried out a Principal Component Analysis (PCA) on the species taxonomic relative abundances in the samples (Figure 4a in [5]), with rough separation observed in the clustering of the four states.

If one looks (Figure 3) at the evolution of the CCD time-series samples using the major microbiome families, the 6th sample seems to move to an ecology more like healthy, in that the family *Bacteroidaceae* is much larger in CCD6 than in any of the other CCD time points.

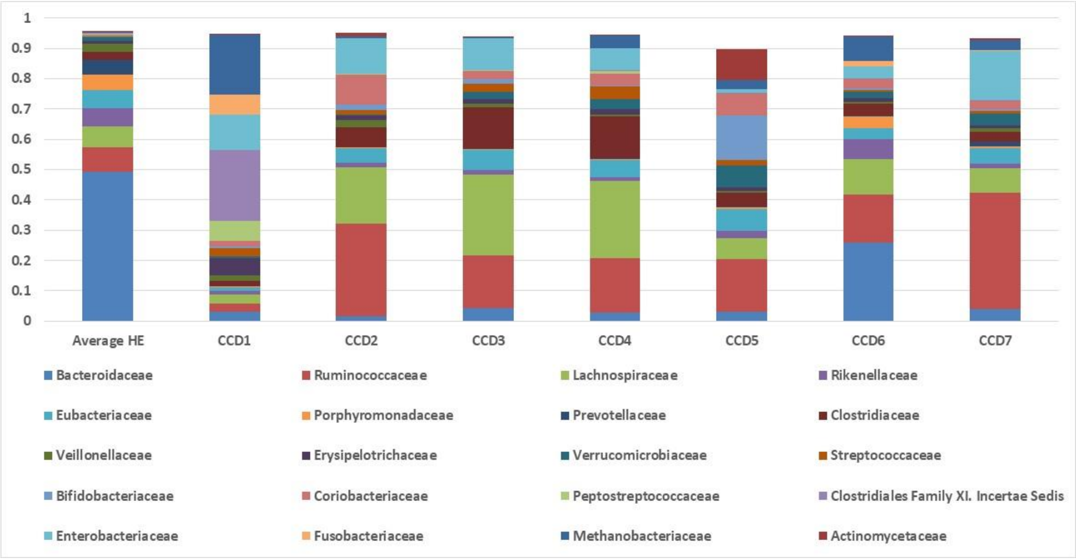


Figure 3: Evolution of the CCD samples microbial ecology compared to the average of the healthy samples (leftmost bar). Stacked bars show relative abundance of microbial families greater than 1% in the samples. The bars add to 100% when all families are included.

From Taxonomy to Function

The question arises as to whether function measured by gut microbiome gene relative abundance might reveal the patterns of disease difference even better than taxonomy. It is well known that in healthy people, even though there is a large variation in the relative abundance of the *Bacteroidetes* to *Firmicutes* Phyla across subjects in the gut microbiome, the variation in function across healthy individuals is quite low (Figure 2 (stool) in [8]).

When the taxonomy of our samples in Table 1 were first computed, it was reported [3] that the relative abundance of the protein families in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [4] was also computed. When we carry out a PCA of the relative abundance of the 10,192 KEGG protein families (Figure 4), we see almost perfect separation of the clusters, much better than was seen in the PCA derived from the relative abundance of the taxonomic species (Figure 4a in [5]).

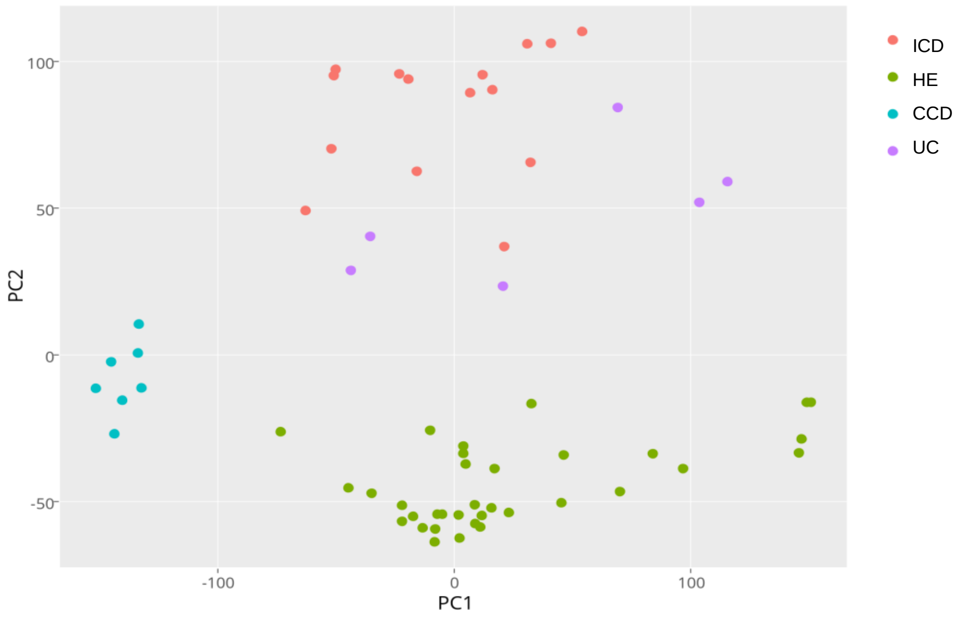


Figure 4: In this PCA of the relative abundance of the 10,192 KEGG protein families across samples, colored by the different subclasses, we see near perfect separation between the different cohorts.

Furthermore, it was discovered [5] (Figure 5) that by using PCI's proprietary machine learning algorithm, there were many KEGG protein families with 1 to 2 orders of magnitude difference in relative abundance between healthy and IBD samples.

Inspired by a quote by E.O. Wilson*:

"The crucial first step to data analytics is subspace selection. If you get to the right subspace, everything else is likely to be easier."

* "The crucial first step to survival in all organisms is habitat selection. If you get to the right place, everything else is likely to be easier" – E.O. Wilson

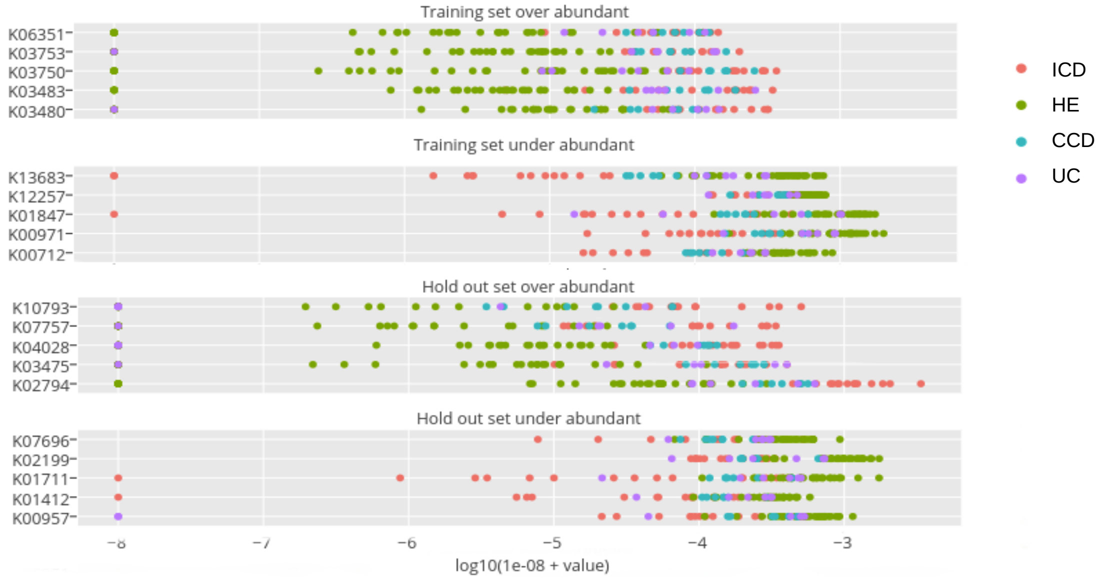


Figure 5: Distribution of the relative abundance of KEGG protein families selected by from use of PCI's machine learning algorithm that discriminate between healthy and disease states. The horizontal axis is the relative abundance values of the KEGG protein families on a logarithmic scale for each of the samples.

In a follow-up paper [9] several statistical tools, including linear regression and topological data analysis (TDA) have been used to show that the three disease subtypes have KEGG protein families that provide clear separation.

The Pattern Computer Approach

Pattern Computer uses a proprietary system to discover new patterns in complex, high-dimensional data sets. Without any prior knowledge, what can we automatically learn from high-dimensional data? If the variables are uncorrelated, then the system is not high-dimensional; instead, it should be viewed as a collection of unrelated univariate systems. If correlations exist, then some common cause or causes must be responsible for generating them. Pattern Computer uses a model-free, mathematically principled approach without prior assumptions to find answers to these questions. We look for latent factors so that, conditioned on these factors, the correlations in the data are minimized — as measured by multivariate mutual information. We look for the simplest explanation that accounts for the most correlations in the data. We illustrate our approach, methodology, and tools to learn more about what a healthy versus unhealthy microbiome states look like and understand more behind the dynamics of the state transitions between healthy and disease states.

The microbiome team made the dataset described above available to the Pattern Computer team to see what additional insights might be generated by use of the Pattern Computing approach. (Note all analysis is done on the $\log_{10}$ transform of the original KEGG data). Using the Pattern Computer toolset, we identified a subspace of 39 out of the 10,192 KEGGS in the original dataset. This subspace captures significant dynamical structure contained within the full data space. To visualize these patterns, we first analyzed the 39-dimensional subspace by computing the cross-correlation between the patients in this subspace. The results from using the Spearman cross-correlation analysis are shown in Figure 6. The clinical subgroups can be clustered from the cross-correlation matrix; clusters denoted by the black boxes. Notice several individuals, yellow boxes, strongly cross correlate with both their own subgroup and with the cluster of healthy patients.

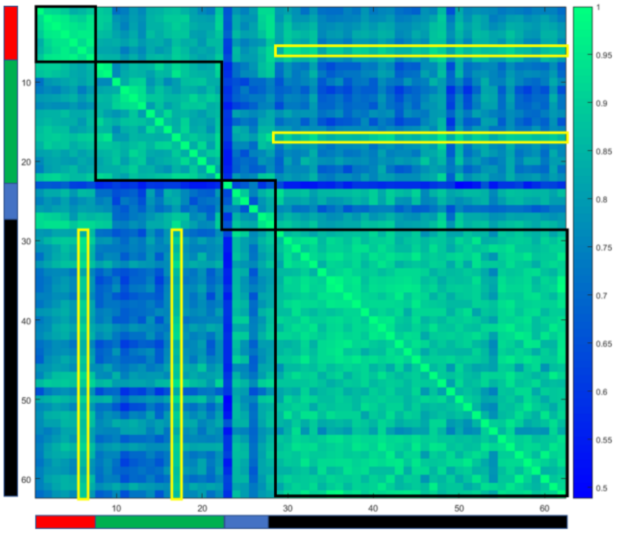


Figure 6: Cross-correlation plot using Spearman correlations of the patients in the 39-dimensional subspace identified by Pattern Computer. The patients are in order with CCD 1-7 (Red), ICD 8-22 (Green), UC 23-28 (Blue), and Healthy 29-62 (Black).

We then utilized two separate embedding techniques to visualize the reduced subspace. Both t-Distributed Stochastic Neighbor Embedding (t-SNE) and principal component analysis were used to embed the 39-dimensional subspace into 3 dimensions for visualization. Because the two methods returned markedly similar embeddings, we refer only to the PCA embedding shown in Figure 7.

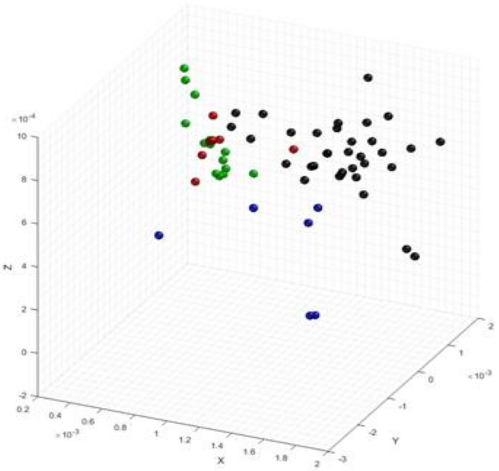


Figure 7: PCA embedding from the 39-dimensional subspace for visualization. The patient clusters: CCD (red), ICD (green), UC (blue), and Healthy (black).

The PCA embedding encapsulates the full structure of the distribution of the data in the 39-dimensional subspace, revealing four sub-clusters representative of the four clinical groups along its three dimensions. In the prior paper (Figure 4a in [5]) using species PCA did not lead to a clear separation between UC (Ulcerative colitis) and HE (healthy) groups. Note that our method verifies the cohort separation that was observed using the KEGG protein families in [5]. It also very interestingly shows a set of individuals that appear to bridge between the cluster of healthy individuals and a super cluster defining the disease states comprised of the CCD and ICD clinical subgroups as depicted in Figure 8. The spatial distribution of clusters highlights key samples which identify with more than one clinical cluster. In other words, the placement of patient samples can denote their similarity to one another.

The results are consistent with the Spearman cross-correlation analysis whereby we see a few samples, strongly cross correlating between the multiple clusters (see Figure 6). Some healthy individuals show certain connections to the disease states as some are clustered closer to the UC or CD subgroups. At the same time, only one of the CCD data points clearly identifies as close to healthy as its own subgroup.

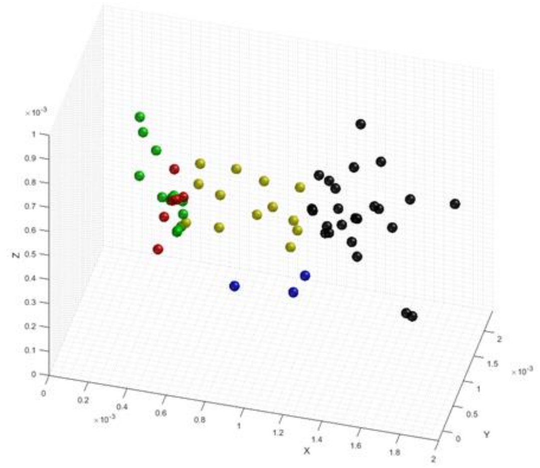


Figure 8: Patients in Yellow identified in the PCA analysis which form a "bridge" between the healthy and disease states.

Figure 9 shows the dynamic transitions of the CCD patient over their 7-time states through the PCA embedding. In the context of the PCA embedding, and assuming we can trade space for time, (in other words, the space in which the data points live describe the dynamics of the individual as they transitioned through time towards different degrees of health), we seek several hypotheses explaining the individuals forming the bridge linking the healthy and disease states as shown in Figure 8.

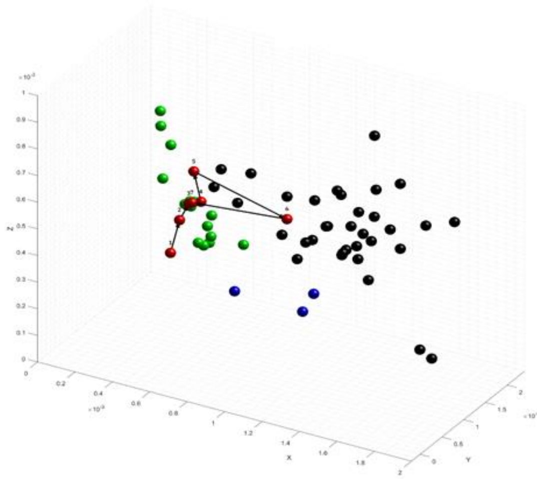


Figure 9: Orbit of the CCD data time-series with respect to the 39-dimensional KEGG subspace.

In Figure 9, CCD6 stands out because it strongly cross correlates between both the CCD cluster and the HE cluster, as indicated by one of the two yellow boxes, through the Spearman cross correlation analysis between individuals, Figure 6, but it still has correlation to its own subgroup. This could be interpreted that CCD6 indicates an in-between disease and healthy state, thus marking part of the transition from healthy to the disease state. Note that this is what one sees in the taxonomic microbial ecology evolution in Figure 3. Identifying 12 patients forming the bridge in Figure 8, we analyzed the spatial distribution of the 39 KEGG protein families individually.

Out of the original 39 KEGGS, nine KEGG protein families showed clear indication of spatial structure not previously investigated: K00330, K00348, K00351, K00434, K00604, K00607, K00609, K00633, and K03671. Prior work proposed a hypothesis that ignored the dynamics of this system in favor of a diagnostic measurement which separated the diseased from the healthy. While these KEGG protein families do show distinct differences in the averages in each subgroup, as validated using the Kolmogorov-Smirnov test, the most prominent and interesting property is that they show a “continuous” change moving from the disease to healthy subgroups within the PCA subspace. To identify this change, we analyzed the distribution of each KEGG relative to a one-dimensional slice through the PCA embedding. This allowed us to identify the KEGGs showing a “continuous” change moving from the healthy to disease state. For example, in Figure 10 for K03671, we compared its distribution along the first component of the PCA embedding where we find a significant linear correlation, $r^2 = 0.82$. We selected the KEGGs with most significant (perhaps non-linear) gradients within the space. These indicated those KEGG protein families that fluctuate across a spectrum of values which seem to span the complex dynamics that lead to an individual’s current condition.

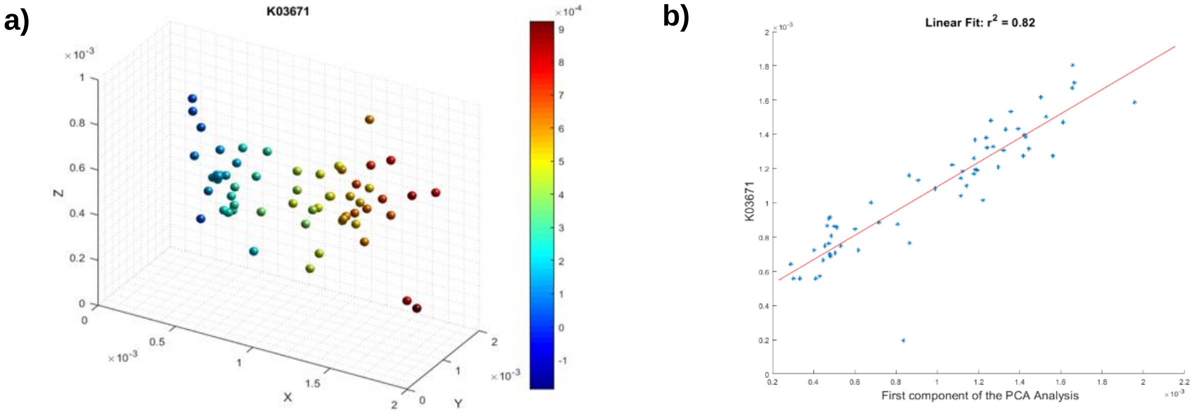


Figure 10: **a)** Visualization of the spatial distribution of K03671 values across patients in the 39-dimensional subspace, **b)** Spatial distribution of K03671 with respect to the first coordinate of the PCA analysis. (Recall all analysis is done on the log10 transform of the original KEGG data).

Our preliminary hypothesize was that nine out of the original 10,192 KEGG protein families have significant dynamical patterns and structure which seem to indicate the transition from disease states to health within IBD. Further research into these nine KEGG protein families revealed that six of the nine KEGG protein families identified are related to oxidative phosphorylation. See Figure 11 for three of these six KEGG protein families. Specifically, K03671 (Figure 11a) and K00640 (Figure 11b) are associated with Combined Oxidative Phosphorylation Deficiency disease (COXPD). K00351 (Figure 11c) is an oxidoreductase enzyme. In contrast, Figure 12 shows an example of two KEGGs drawn at random that clearly do not show spatial structure with respect to their distribution amongst the patients.

With these analyses, we believe that IBD, like other inflammatory diseases, may be associated with abnormal oxidative phosphorylation or oxidative stress [10]. Oxidative phosphorylation produces reactive oxygen species (ROS) in both prokaryotes and in the mitochondria of eukaryotes. Microbial ROS production affects the innate immune response, influencing the integrity of the intestinal epithelial barrier (2) which is compromised in IBD. Future research is required to more definitively follow up this finding.

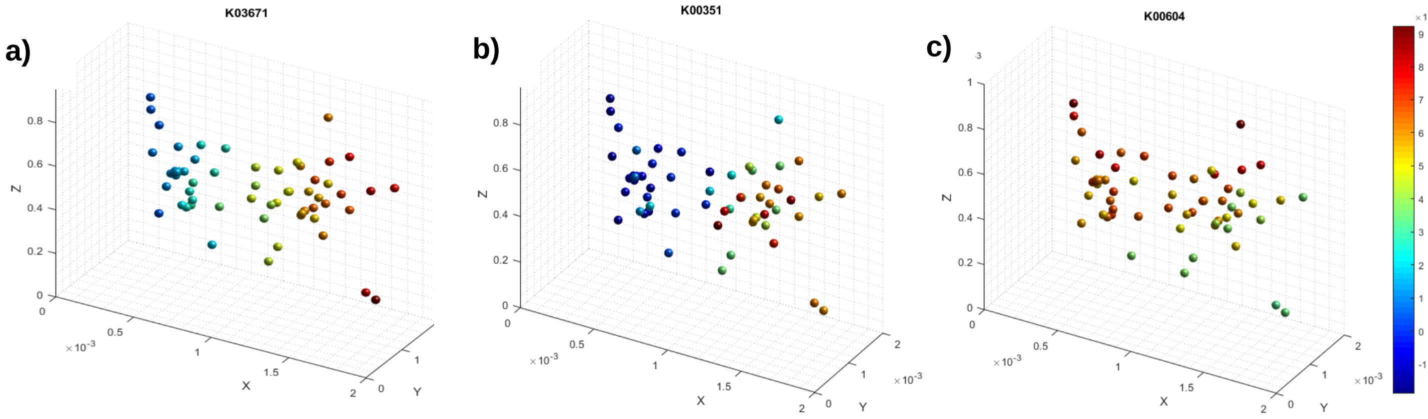


Figure 11: Spatial distribution of KEGG values across patients in the 39-dimensional subspace a) K03671, b) K0351 and c) K00604.

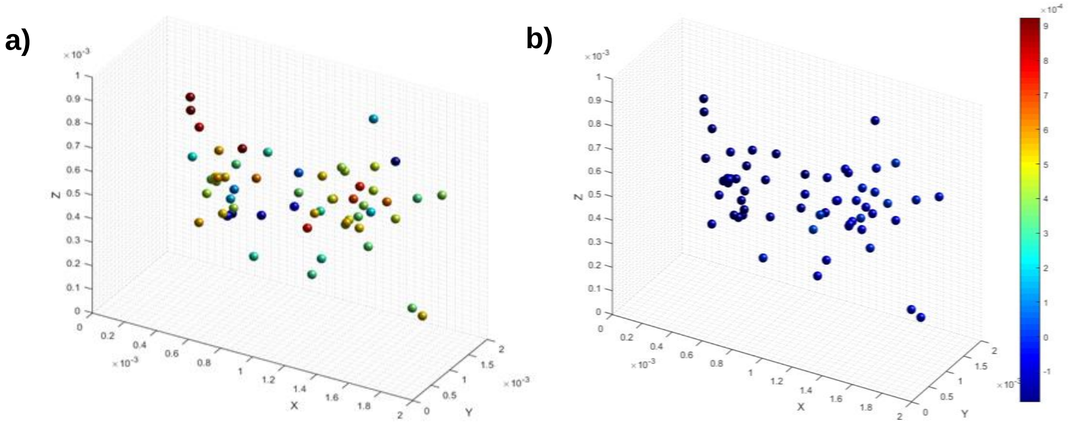


Figure 12: In contrast to Figure 11, **a)** KEGG whose distribution has no coherent spatial structure across the patients, **b)** KEGG with a near constant spatial distribution.

Presently, the number of features measured within many biological datasets (omics data, in particular) and thus, the potential combinations of subspaces to be explored are daunting. If science is to progress, its ability to create new hypotheses based upon data insights is essential. Thus, the scientific community requires a new generation of data mining tools that can find intrinsic patterns within subspaces of a dataset worth investigating in an automated fashion. In this white paper, PCI's tools identified a subspace of 39 KEGG protein families out of 10,192, or one single 39-dimensional subspace out of approximately 10^{110} combinations of 39-dimensional subspace in under a few hours of run time on our machines, capturing significant intrinsic structure allowing for new hypotheses about IBD to be formulated.

References

1. J. Gilbert, R. A. Quinn, J. Debelius, Z.Z. Xu, J. Morton, N. Garg, J. Jansson, P. C. Dorrestein, and R. Knight, “Microbiome-Wide Association Studies Link Dynamic Microbial Consortia to Disease.” *Nature*, vol. 535, pp. 94–103, 2016.
2. J. Halfvarson, C. J. Brislawn, R. Lamendella, Y. Vázquez-Baeza, W. A. Walters, L. M. Bramer, M. D’Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E. E. McClure, M. F. Dunklebarger, R. Knight, J. K. Jansson, “Dynamics of the human gut microbiome in Inflammatory Bowel Disease,” Nat. Microbiol. Vol. 2 p. 17004, 2017.
3. 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, p. 25, 2013.
4. 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, pp. D457–D462, 2016; M. Kanehisa and S. Goto, “Kegg: kyoto encyclopedia of genes and genomes,” Nucleic Acids Research, vol. 28, pp. 27–30, 2000.
5. M. Yazdani, B. C. Taylor, J. W. Debelius, W. Li, R. Knight, and L. Smarr, “Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease” *Proceedings of the 2016 IEEE International Conference on Big Data*, 2017.
6. Cleynen, et al., “Inherited determinants of Crohn’s disease and ulcerative colitis phenotypes: a genetic association study,” *The Lancet*, vol. 387, pp. 156–167, 2016.
7. J. Jansson, B. Willing , M. Lucio , A. Fekete, J. Dicksved, J. Halfvarson, C. Tysk, and Philippe Schmitt-Kopplin, "Metabolomics Reveals Metabolic Biomarkers of Crohn's Disease," *PLOS ONE*, vol. 4, e6386, 2009; B. P. Willing, J. Dicksved, J. Halfvarson, A. F. Andersson, M. Lucio, Z. Zheng, G. Jarnerot, 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, pp. 1844–1854, 2010.
8. H. M. P. Consortium et al., “Structure, function and diversity of the healthy human microbiome,” *Nature*, vol. 486, pp. 207–214, 2012.
9. L. Smarr, M. Yazdani, et al, in preparation.
10. L. E. Glover, S. P. Colgan, “Hypoxia and metabolic factors that influence inflammatory bowel disease pathogenesis,” *Gastroenterology,* vol. 140, pps.1748-55, 2011.; R. M. Jones, J. W. Mercante, A. S. Neish, “Reactive oxygen production induced by the gut microbiota: pharmacotherapeutic implications,” Curr. Med. Chem. Vol.19, pps.1519-29, 2012.