CS 109B/Stat 121B/AC 209B/CSCI 109B: Final Project
Glickman, Protopapas

Module 2: Microbiome

Faculty

Georg Gerber, MD, PhD, MPH
Assistant Professor, Harvard Medical School
Member of the Harvard-MIT Health Sciences & Technology Faculty
Chief, Division of Computational Pathology
Co-Director, Massachusetts Host-Microbiome Center
Department of Pathology, Brigham and Women's Hospital

Dr. Gerber is a computer scientist, microbiologist and physician who heads a research lab focused on creating novel machine learning methods and high-throughput experimental systems to understand the role of the microbiota in human diseases and applying these findings to develop new diagnostic tests and therapeutic interventions to improve patient care. He also co-directs the MA Host-Microbiome Center, which provides extensive microbiology, molecular, and animal model resources to microbiome researchers nationally. Dr. Gerber holds a PhD in Computer Science from MIT (statistical machine learning) and an MD from Harvard Medical School.

Travis Gibson, PhD
Instructor, Harvard Medical School
Department of Pathology, Brigham and Women's Hospital

Dr. Gibson works at the intersection of systems, control, and machine learning. His PhD was in adaptive (learning) control from MIT, and postdoctoral training is in the area of statistical inference and microbial dynamics. His work has been published in high-impact journals as well as in flagship control theory and machine learning venues. Dr. Gibson also has experience developing safety critical algorithms that are currently flying on Boeing and NASA experimental aircraft.

Problem statement

Through this module, you will have an opportunity to learn about the microbiome, one of the most exciting research fields in science and medicine today. You will also gain experience with analyzing complex time-series data and familiarity with Bayesian nonparametric methods, techniques that are also very useful in other domains beyond the microbiome.

Your body has as many microbial cells as human cells. These trillions of organisms, deemed the human microbiome, form ecosystems throughout your body that are very important for your health. Until quite recently, little was known about the microbiome. With the advent of new

sequencing and other high-throughput experimental techniques, it is now possible to measure the composition and activities of these complex and fascinating ecosystems in great detail.

There is mounting evidence that the microbiome is critical in many aspects of health human including function of the digestive tract, brain, skin and reproductive and immune systems. Conversely, disruption of the microbiome has been implicated in a variety of major human diseases including infections, arthritis, food allergy, cancer, inflammatory bowel disease, neurological diseases, and obesity/diabetes. This evidence has sparked intensive research and development among biotechnology and pharmaceutical companies to develop approaches to manipulate the microbiome for therapeutic benefit. These approaches include bacteriotherapies, or “bugs as drugs,” cocktails of multiple bacterial strains designed to be given to patients as living therapeutics. There have been >45 disclosed equity investments in microbiome therapeutics since 2010, totaling > $840M, with a projected market size of $3.2B by 2024.

The human microbiome is inherently dynamic, due to interactions among microbes, with our bodies, and with the environment. These complex dynamics begin at birth, as we are colonized with microbes, and continue in the healthy adult as microbial populations vary with diet and a myriad of other host and environmental factors. At any point in time, the microbiome can be dramatically altered, either transiently or long term, by diseases such as infections or medical interventions such as antibiotics. Analyzing these dynamics is critically important for understanding the role of the human microbiome in health and disease, and for designing therapeutic interventions to alter the microbiota.

Microbiome time-series data is very complex and high-dimensional, measuring the population dynamics of hundreds to thousands of microbial taxa. This necessitates sophisticated and creative computational techniques for analysis. In this module, you will have the opportunity to learn about the microbiome and its role in human disease and apply advanced machine learning/statistical methods to datasets to solve challenging problems in the microbiome field. In particular, you will have access to published and new datasets with dense time-series measurements of the microbiome subjected to perturbations including changes in diet, antibiotic exposure, and infection. Your challenge will be to infer latent interaction structure in the microbiome and forecast the effects of perturbations on the state of the microbiome. Solving this challenge has important practical applications in rationally designing bacteriotherapies and predicting what will happen when we administer them to patients.

Lectures

Lecture #1: Microbiome science boot camp. This lecture will cover critical domain knowledge including basics of microbiology, gastrointestinal anatomy/physiology, and aspects of microbiome therapeutic development.

Lecture #2: Microbiome bioinformatics and mathematical models. This lecture will build understanding of: (a) bioinformatics methods and tools used to transform raw next generation sequencing data into meaningful microbiome measurements, and (b) dynamical systems models conventionally used for understanding behavior of the microbiome as an ecosystem.

Lecture #3: Statistical and machine learning approaches for microbiome time-series analysis. This lecture will introduce cutting edge approaches for analyzing the microbiome over time, including nonparametric Bayesian methods. Our emphasis will be on principled inference from microbiome time-series data given its special properties.

Datasets for analysis

1. Published data of a defined microbiome exposed to infection with the pathogen *Clostridioides difficile*, consisting of 26 time-points and millions of sequencing reads per time-point. One objective with this dataset is to infer latent interaction structure among microbes, to develop a therapy to protect against *C. difficile* infection.

2. Published data of a defined microbiome exposed to different diets, consisting of 50+ time-points and millions of sequencing reads per time-point. One objective with this dataset is to forecast the effects of the different perturbations on the microbiome.

3. New data of uncharacterized human microbes exposed to different diets and antibiotics, consisting of 50+ time-points and millions of sequencing reads per time-point. This is the largest such study to date to our knowledge. One objective with this dataset is to forecast the effects of the different perturbations on the microbiome. A major challenge will be to scale inference methods to the full human microbiome.