Statistical approaches for studies of human microbiome

Dr. Siddhartha Mandal

Abstract:

Human microbiome plays a crucial role in health and diseases, with recent research unravelling the myriad ways in which these associations are manifested. Microbial communities associated with the human body sites are complex ecosystems, with large number of interactions and functions. While traditional microbiology focused on microbial pathogens and few beneficial bacteria; advances in high-throughput technologies have presented evidence of a broader spectrum of microbial functions in relation to chronic diseases, brain function and neurodevelopmental outcomes. Understanding factors regulating our microbiota and the impact of microbiota on health requires development of appropriate statistical methodology while acknowledging the constraints of the data.

In this presentation, we shall start with an introduction to microbiome, highlighting its implications on human health to motivate the research questions. Few major questions of interest in studies of microbiome are *"How do microbial communities cluster between different groups, for example, babies born by C-section and vaginal delivery?"* or *"Which microbial taxa are differentially abundant between mode of delivery?"*. Data generated from microbial surveys are relative abundances of microbial taxa (which belong to a phylogenetic tree), while the actual abundances are unobservable quantities. This results in compositional data, where observations on each subject are multivariate vectors belonging to a simplex. Application of standard statistical methodologies on such data, to make inferences on the abundances, often lead to false interpretations. Existing approaches to detect differentially abundant microbes either discount the underlying compositional structure in the microbiome data or use inappropriate probability distributions that may potentially increase false discoveries. We introduced a statistical framework called Analysis of Composition of Microbiomes (ANCOM) that accounts for the underlying compositional structure in the data, and, unlike existing approaches, can compare the abundances of microbiomes at the ecosystem level across populations. ANCOM makes no distributional assumptions, and recent modifications allow inclusion of confounders and longitudinal study designs. In addition, ANCOM scales well to compare microbial communities involving thousands of taxa. We shall illustrate some aspects of the methodology using publicly available human gut microbial datasets. In addition, we shall explore some of the potential future methodological and applied research problems in this area and how these may be relevant in the case of Indian populations and related health problems.