A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links

We recently presented a three-pronged association study that integrated human intestinal microbiome data derived from shotgun-based sequencing with untargeted serum metabolome data and measures of host physiology. Metabolome and microbiome data are high dimensional, posing a major challenge for data integration. Here, we present a step-by-step computational protocol that details and discusses the dimensionality-reduction techniques used and methods for subsequent integration and interpretation of such heterogeneous types of data. Dimensionality reduction was achieved through a combination of data normalization approaches, binning of co-abundant genes and metabolites, and integration of prior biological knowledge. The use of prior knowledge to overcome functional redundancy across microbiome species is one central advance of our method over available alternative approaches. Applying this framework, other investigators can integrate various '-omics' readouts with variables of host physiology or any other phenotype of interest (e.g., connecting host and microbiome readouts to disease severity or treatment outcome in a clinical cohort) in a three-pronged association analysis to identify potential mechanistic links to be tested in experimental settings. Although we originally developed the framework for a human metabolome-microbiome study, it is generalizable to other organisms and environmental metagenomes, as well as to studies including other -omics domains such as transcriptomics and proteomics. The provided R code runs in ~1 h on a standard PC.