A Physics-based Dimensionality Reduction Approach for Biological Data

Prof Purushottam Dixit

Department of Physics, University of Florida

Abstract

Advances in omics technologies have allowed a high dimensional characterization of biological systems. However, due to complexity of interactions, bottom-up mechanistic models are not feasible. We need a top-down methods to integrate available data and mechanistic information. The maximum entropy (Max Ent) principle has proved to be a promising framework for such integration. However, there are several limitations to Max Ent. The most important being that the modeler is required to a priori identify appropriate constraints. Max Ent models are notoriously difficult to learn, prohibiting their applications to problems with large data dimensions. To address these issues, we propose the agnostic maximum entropy superstatistics (AMES). AMES has several salient features. First, in AMES, the modeler only specifies the total number of constraints. The constraints are optimally learned from the data and a Max Ent model is fit to those constraints. Second, because of this optimal choice, AMES inference is significantly faster than typical Max Ent, allowing us to analyze very high dimensional data (dimensions) that remain well out of the reach of current Max Ent methods. Third, AMES is a non-linear latent-space based dimensionality reduction method. We will develop AMES and discuss in detail two specific applications (1) predictive models of host-microbiome interactions and (2) models of protein sequence variation.