Research in molecular biology and other fields often requires the analysis of datasets in which multiple related sources of data are available for a common sample set. We describe two exploratory methods for the integrated analysis of such datasets: Joint and Individual Variation Explained (JIVE) and Bayesian Consensus Clustering (BCC). JIVE gives a general decomposition of variation consisting of three terms: a low-rank approximation capturing joint variation across data sources, low-rank approximations capturing structured variation individual to each data source, and residual noise. JIVE quantifies the amount of joint variation between data sources, reduces the dimensionality of the data in an insightful way, and allows for the visual exploration of joint and individual structure. BCC is a tool to cluster a set of objects based on multi-source data. The Bayesian model permits a separate clustering of the objects for each data source that adhere loosely to an overall clustering. We illustrate the above methods with applications to publicly available data from The Cancer Genome Atlas. This is joint work with collaborators at The University of North Carolina and Duke University.

Abstract