THE UNIVERSITY OF CHICAGO

Department of Statistics

DISSERTATION PRESENTATION AND DEFENSE

KUSHAL DEY

Department of Statistics
The University of Chicago

Model Based Visualization of Structure in Biological Data

THURSDAY, June 7, 2018, at 2:00 PM
Jones 226, 5747 S. Ellis Avenue

ABSTRACT

Biological Data comes in varied forms and the scale of the data is typically large, which often necessitates distinct modeling frameworks and tools to process, analyze and visually summarize the data. An overarching theme of this doctoral thesis is to suggest such novel model based visualization tools for different biological problems.

The second chapter of this thesis extends the concept of a mixed membership model, popularly known as ADMIXTURE model in population genetics and topic model in Natural Language Processing (NLP), to the context of RNA-sequencing read expression data in genetics. Applied to data from the GTEx project on 53 human tissues, this approach highlights similarities among biologically-related tissues and identifies distinctively-expressed genes that recapitulate known biology. Applied to single-cell expression data from mouse preimplantation embryos, this approach highlights both discrete and continuous variation through early embryonic development stages, and identifies genes involved in a variety of relevant processes from germ cell development.

The third chapter extends similar mixed membership models to analyzing DNA damage patterns in ancient DNA (aDNA) samples, and explore and jointly summarize multiple aDNA samples together with modern samples. Applied to a combined data of modern and ancient individuals from multiple studies, this approach clearly distinguished moderns and ancients irrespective of DNA extraction, lab and sequencing protocols. Additionally, we found that the grades of membership from the fitted mixed membership models can be reflective of relative levels of contamination in the data.

The visual summary of DNA damage patterns, depicted above, includes a version of logo plot that highlights enrichment and depletion of damage features with respect to a background level of mismatch features computed from modern individuals. We call this representation the *Enrichment Depletion Logo (EDLogo)* plot and present a comprehensive overview of this logo plot in the fourth chapter. We also propose an extension of typically character driven logo plots to string based logos.

In the fifth chapter, we propose an adaptive method for shrinking correlation matrices that leads to a parsimonious representation of the underlying association structure between variables. This method is flexible in handling data matrices with missing observations and accounts for the differences in the number of samples with non-missing observations for a pair of variables in a model based way. Even with no missing data, under small n, large p settings, this method outperforms other popular approaches to correlation shrinkage and is flexible enough to extend to other correlation-like quantities such as the word-word cosine similarity values from *word2vec* models in Natural Language Processing (NLP).

For information about building access for persons with disabilities, please contact Laura Rigazzi at 773.702-0541 or send an email to lrigazzi@galton.uchicago.edu. If you wish to subscribe to our email list, please visit the following web site: https://lists.uchicago.edu/web/subscribe/statseminars.