Matrix Visualization and Information Mining for Genomics/Proteomics Data Structure

Chun-Houh Chen

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC

Abstract

In genomics and proteomics studies, phenotype and genotype data are often collected and presented as raw data matrices and proximity matrices. Many statistical techniques, particularly multivariate methodologies, focus on extracting information from these matrices. Rather than rely solely on numerical characteristics, matrix visualization allows one to graphically reveal structure in a matrix.

Some of these phenotype and genotype data are of continuous nature (gene expression profiles, metabolite profiles for example) while others may contain ordinal information (short tandem repeat (STRP), mouse mutagenesis database) or categorical phenomenon (pathway database, phylogenetic profiles such as clusters of orthologous groups of proteins (COGs), Single nucleotide polymorphisms (SNPs)). Conventional matrix visualization tools such as Cluster/TreeView can be employed for studying data with continuous structure but are insufficient for exploring more complicated information structures in the statistical modeling of longitudinal, categorical, dependent or other complex data.

In this presentation, we will first briefly review conventional matrix visualization packages, then provide a more detailed description of its general framework, along with some extensions. Possible research directions in matrix visualization and information mining for phenotype and genotype data structures are sketched.