

Wide-scale Comparison of Transcriptome Data

Raymond Lim (raylim@mm.st)^a, Paul Pavlidis^a

A common problem in genomics experiments is interpreting a resulting "hit list" list of interesting genes. Here we present work on an approach for summarizing and exploring "hit lists" that makes use of the large amount of gene expression data in public repositories such as the Gene Expression Omnibus. We compare the query list with data sets that we have analyzed for differential expression of genes. Studies that have similarities to the given hit list yield additional insights, help contextualize studies, and serve as a basis for future meta-analysis. A conceptually similar problem that we address is the classification or clustering of datasets. Both problems require a method for determining distances between data sets based on their differential expression. We tested and benchmark several methods using manually annotated datasets. The method that performs best according to our evaluation process is based on Kendall's Tau top-k distance. We demonstrate the effectiveness of this approach in a few case studies.

^a Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada