Erik Taylor, Biomedical Engineering

University: University of Texas
Year in School: Junior
Hometown: Austin, Texas
Faculty Mentor: Dr. Dong Xu, Computer Science
Funding Source: NSF-REU Program in Biosystems Modeling and Analysis

The automation of microarray data analysis to ameliorate biochemical pathways

Erik Taylor and Dong Xu

The goal of this study is the annotation of a gene cluster in microarray data with probable biochemical pathways. In our experiment each gene from Arabidopsis microarray data was compared to each gene from Arabidopsis KEGG pathway, and a similarity calculation was made. For the 180 unique pathways for Arabidopsis in KEGG, the maximum similarity was annotated based on which gene with KEGG pathway annotation the microarray gene was most similar to. A single gene will likely have many GO ID terms, creating a good thoroughfare for calculation of gene similarity. When two GO IDs were compared a number that is based on parent GO terms between zero and one was assigned. Each term to term similarity was summed, and divided by the number of comparisons, giving a similarity rating between zero and one for two genes. The genes are next grouped using fuzzy c-means clustering, and each cluster is annotated based on maximum pathway membership. Two different matrixes were constructed for data analysis: Sim contained KEGG genes matched at similarity 1 and all other matches with a number between zero and one, called fuzzy matches generated by GO ID similarity described above, and Sim4 contained all fuzzy matches. Analysis of both matrixes for KEGG genes in properly labeled pathways was done at 19, 22, and 40 clusters resulting in Sim with 73, 82, and 97 percent matched respectively and Sim4 with 24, 32, and 51 percent matched respectively. Analysis of each cluster is done using a similarity sum for each pathway to find a maximum, and division by the number of genes in the cluster. Validity of this is found using the Sim matrix, and the analysis of clusters that only contain KEGG genes, where a return of 1 is found. Other clusters with significant values were found, often containing only unknown genes.