IDENTIFICATION OF CELL FUNCTION PATHWAYS OF GENES FROM 60 HUMAN CELL LINES USING NEURAL NETWORKS AND RANDOM FORESTS

Melanie Smith

Bioengineering and Bioinformatics Summer Institute
Virginia Commonwealth University
Richmond, VA 23284

Introduction

Using microarray data, the effects of disease, mutations, and drugs on the genomic level can be discovered. By analyzing the data with cluster analysis, patterns in gene expression can be used to describe and classify disruptions of normal cell function. In order to understand what the changes in gene expression mean in terms of what functions are being affected in the cell, genes that are associated with these diseases, mutations, and drugs can be identified and their functions can be characterized.

Using different computer models, the primary, secondary, tertiary, and possibly quaternary functions of genes can be determined based on cluster analysis of microarray data. Once the functions of genes are determined, groups of functions could be connected if the same genes are classified together in their primary and secondary functions as well as other functions. For instance, it is expected that the TCA cycle and glycolysis are closely linked and many of the genes in these pathways will be in both.

Two computer models were used to determine the functions of over 6,000 human genes from 60 cell lines. The cell lines represent different types of cells such as liver, blood, and skin cells. The genes were categorized into 21 different cell pathways defined by the *Kyoto Encyclopedia of Genes and Genomes*.

Neural Networks is a computer model that assigns a probability for each pathway that the gene in question could be a part of. The pathway with the highest probability for that gene is classified as the primary pathway of that gene. The second probability becomes the secondary pathway and so on.

Random Forest is another computer model that was used to classify the genes. For this model, the gene data was run through the model 500 times and each time the model guesses one pathway that the gene is in. After 500 votes, the pathway with the largest amount of votes is determined to be the primary pathway of the gene, the second largest amount of votes the secondary pathway, and so on.

Once all of the genes were classified into their probable pathways, genes that had defined functions in the *Kyoto Encyclopedia of Genes and Genomes* were analyzed for how accurately the neural network correctly classified the primary pathway.

The next step in the research is to define the secondary, tertiary, and quaternary pathways of the genes and thereafter find correlations between the pathways. Then, a network of connected pathways will be assembled and compared to the current understanding of how cell functions are related. The network will be created in a database.

Methods

The microarray data of the 6,000 human genes have been analyzed by Neural Networks and Random Forest models and the data have been placed into Microsoft Excel. Using Excel, the data will be analyzed to determine the pathways of the genes and how they correlate. For each gene, a relative probability will be calculated for how likely it is that the gene is in any of the 21 pathways. A threshold of relative probability will be defined for all of the pathways to potentially enrich for genes with the highest confidence of classification. Associations will be determined between primary, secondary, tertiary, and quaternary pathway classifications and analyzed for statistical significance. Those associations with significance will be further analyzed and in some cases validated using previously published studies.

Possible Results and Conclusions

Once the pathways are identified and correlations are identified, the information will be compared to known data about the correlation of the genes and the pathways. Once these can be verified, a graphic network will be created and the relationships will be available through a database.

Identifying the pathways of unknown genes is important when analyzing what genes are being disrupted by different stimulus such as disease. Knowing what pathways are disrupted can aid in drug design and deciding how harmful or beneficial a drug can be.

The elucidation of connected pathways can provide insight into how a drug targeting a gene that affects the cell cycle is also disrupting a secondary pathway of that gene that is essential to a cell's proper function.

The methods used in this research can be used to model other organisms' gene functions as well and can identify the metabolic pathway associations.

REFERENCES

Breiman, Leo and Adele Cutler. Random Forests. http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (8 June 2006).

Gene Ontology. (2006). http://www.geneontology.org/index.shtml (8 June 2006).

Hughes, Timothy R., Mathew J. Marton, Allan R. Jones, Christopher J. Roberts, Roland Stoughton, Christopher D. Armour, Holly A. Bennett, Ernest Coffey, Hongyue Dai, Yudong D. He, Mathew J. Kidd, Amy M. King, Michael R. Meyer, David Slade, Pek Y. Lum, Sergey B. Stepaniants, Daniel D. Shoemaker, Daniel Gachotte, Kalpana Chakraburtty, Julian Simon, Martin Bard, and Stephen H. Friend. 2000. Functional discovery via a compendium of expression profiles. Cell. 102: 109-126.

Ko, Daijin, Wanyan Xu, and Brad Windle. 2005. Gene function classification using NCI-60 cell line gene expression profiles. Computational biology and chemistry. 29: 412-419.

Kyoto encyclopedia of genes and genomes. (2006). http://www.genome.ad.jp/kegg/ (8 June 2006).