CANCER STUDIES USING BIOINFORMATIC TECHNIQUES

Melanie Smith

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

Introduction

The study of cancer is so diverse that there are many bioinformatic applications that would be useful to evolve the field. Over last summer, I began to improve and quicken the anticancer drug-screening process. Anticancer drugs have been made to target specific genes or their products in order to produce some desired outcome, such as destroying cancerous cells. Using NCI's 60 cell lines, microarray data has been able to reveal whether an anti-cancer drug is actually accomplishing its goal. Researchers have compiled expression profiles for over 100,000 compounds using the 60 cell lines.

In order to better understand how these compounds are affecting the cells, computer models have been used to classify genes into what biological function they perform. Taking this idea one step further, my research goal was to determine what secondary, tertiary, and quaternary functions these genes have. Also, I analyzed the genes to determine if a significant number of genes from a primary function also participated in the same secondary function. By discovering what cell pathways are associated with each other, the effects of drugs can be better understood.

As important as it is to understand how cures for cancer affect the cell, it is also a crucial step in research to understand how causes for cancer affect the cell. During the school year, I will begin to study the Human Papilloma Virus, which is a known causative agent of cervical cancer. Understanding the mechanisms of infection will give greater insight into the parasites ability to cause cancer in its host. In order to address this question on a genomic level, bioinformatic techniques will be used.

Progress Report

Over this past summer, I used Neural Network and Random Forest models to discover what cell functions are related based on gene expression profiles of NCI's 60 cell lines. The tumor cell lines are from lung, colon, breast, prostate, kidney, melanoma, leukemia, central nervous system, and ovary cancers.

The models were created to classify genes based on their function. In order for the models to correctly classify a gene, they needed a large number of references. For the 6,165 genes that I used, 367 of the genes were used as a reference. These genes are well known and their functions were determined by the Kyoto Encyclopedia of Genes and Genomes website. Each gene's expression profile was compared to the reference set, and classified into the pathway that best fit. The 367 training genes only represented 21 pathways of the 300 pathways from KEGG, so the models could only classify genes into these 21 pathways. These 21 pathways are listed in Table one on Page 3. The 21 functions of the cell were chosen from metabolic, genetic information, environmental information, and cellular processes that are well known. Also, a human disease, Huntington's Chorea, was included among the cell functions in order to test the models' ability to classify genes involved in the disease as well as relate pathways to the disease. This could reveal useful applications for other diseases as well.

Table one- The 21 Pathways from KEGG:

- Metabolism
  - Arginine and Proline
  - N-glycans
  - Glycolysis
  - Oxidative Phosphorylation
  - TCA Cycle
  - Pyrimidine
  - Glutathione
  - Valine, Leucine, Isoleucine
  - Porphyrin
  - Purine
  - Glycerolipids
- Genetic Information Processing
  - Ribosome
  - Polymerases
  - tRNA
  - Proteasome
- Cellular Processes
  - Cell Cycle
  - Apoptosis
  - Cell Adhesion
- Environmental Information Processing
  - MAP Kinase
  - PI Kinase
- Human Diseases
  - Huntington's Disease

The neural network model determines how probable it is that a gene is in a pathway and assigns a relative probability for all 21 pathways. For the Random Forest Model, a gene is classified into one pathway, but each time the model is run, it may classify the gene differently. The pathway the gene is classified into most of the time is most likely to be the pathway it belongs to. So the model was run 500 times, the number of times the gene was classified into each pathway was recorded, and the "votes" were converted to relative probabilities (votes/500). This process was repeated 100 times to get a more representative answer. All of the probabilities from both models were analyzed using Microsoft Excel and can be obtained from the Pathway Probabilities link on my web page.

In order to decide what pathways were linked, I wrote a couple of programs in Microsoft Visual Basic. The first program counted how many genes were in a primary pathway and how many of those genes went to the same secondary, tertiary, or quaternary pathway. The second program calculated the binomial distribution to determine how probable it is that the two pathways are related. The number of successes was how many genes from the primary pathway were in the secondary pathway. The number of trials was how many genes in total were classified to the secondary pathway. The probability of success was determined by dividing the number of genes in the primary pathway by the total number of genes in the model. These values were used as p-values and were converted to –log (p-values) to be more easily read. The Pathway Association Scores link on my web page shows the p-values for both models.

Using the pathway association scores, a network of how the different pathways are connected was constructed for both models and for primary versus secondary, tertiary, and quaternary. These diagrams were created in Microsoft Excel and can be viewed by going to the Pathway Networks link on my web page.

Goals for Academic Year

During the academic year, I plan to improve upon my computer programming skills over winter break. Until this summer, I had not experienced any computer programming, and I am interested in learning more so that I can more effectively conduct research next summer. During the summer of 2007, I will be writing programs to analyze anticancer drug expression profiles from NCI's 60 cell lines. So, more experience in computer code will be beneficial to my research.

Along with learning more about computer programming, I will be learning more computer techniques while studying HPV. By studying HPV, I will be discovering more about how cancer occurs from viral infection.

Plan for Academic Year

I have divided my time so that while I am at school, I will be researching HPV; and while I am on winter break, I will be learning computer programming.

Computation

Although my project at school is not directly related to my BBSI project, it is still related to cancer, and any information produced about HPV would be helpful in determining treatment for the cancer it can produce. I will be using various methods to analyze the HPV genome. And, most likely, I will be using the computational tools listed on my home mentor's website called Useful Links that leads to many bioinformatic tools. More specifically, I will likely be using NCBI's Genbank, BLAST, Genome Site, and Gene Expression Omnibus sites.

Courses

As far as the computer programming goes, I plan on learning the C++ programming language. My school doesn't offer it over winter break, but perhaps a community college does, otherwise, I will be on my own. During the summer, I not only programmed in Microsoft Visual Basic but began to write scripts in FileMaker Developer as well; so learning a programming language is not my plan, but I am working toward learning the basic structure and logic of computer programming. I cannot say what project I will specifically be working on next summer, so I cannot explicitly link my research to learning C++. But, I may try analyzing biological data similar to what I've done using what data is available to me. Maybe I will get a better idea of what I am researching as it gets closer to summer, and I can begin applying what I've learned.

Budget

So far, my plans are for free.

References

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

Collins, Jerry M. Ph.D., Associate Director. Developmental Therapeutics Program NCI / NIH. (2006). http://dtp.nci.nih.gov./index.html. (4 Aug 2006).

Developmental Therapeutics Program NCI/NIH. (2006). http://dtp.nci.nih.gov./index.html. (8 June 2006).

D. T. Ross et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics. 2000 March, 24 (3): 225-234.

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

Hughes, Timothy R., et al. 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).