Introduction

The serum proteome has the potential to reveal pathologic processes at the earliest, most curable stage in various cancers. In the case of prostate cancer, a recent study of total protein abundance revealed distinct protein patterns in patients with tumors vs. patients with benign disease.\textsuperscript{[1]} The analysis was limited to these two variables and required several days of supercomputer processing. While this method was able to predict advanced cancer with 95\% accuracy, it could predict benign prostate disease with only \~78\% accuracy and early prostate cancer with only 71\% accuracy. Therefore, we decided to investigate alternative methods for quickly differentiating protein patterns in multiple disease and disease-free states without the limitations of 2-variable analysis or the need for large computing resources.

Analytical Methods

In our pilot study, data files from protein surface-enhanced laser desorption ionization time-of-flight (SELDI-TOF) mass spectroscopy were directly imported into GeneSpring. Using $M/Z$ ratios as unique identifiers, peak height as intensity, serum samples were used to build a GeneSpring experiment. Our data included 15,000 peaks in the 43 known samples from patients with prostate cancer, benign disease, and normal prostate tissue. Our first goal was to eliminate peaks with uninteresting patterns. We used several filtering tools, including k-means clustering, to remove these non-interesting peaks. In Figure 1 10,566 peaks were left unclassified after k-means clustering and were removed from subsequent analysis. Hierarchical clustering (Figure 2) was used to determine the relationship of samples to each other to ensure that there were three predicted classes. To develop a set of predictor peaks, we applied Fisher’s exact test, and identified 20 peaks useful for segregating 17 patient samples taken from the original dataset. Training data was also cross-validated using a k-nearest neighbor algorithm with 5 nearest neighbors used for voting.
Results and Discussion

In the Tree View shown in Figure 2, red branches represent cancer, green branches represent benign disease, and white branches represent disease-free patients. Benign disease appears as a totally separate tree on the right of the graph. It can be seen from the tree that the number of predicted classes in the cancer and normal samples is higher than expected, but overall the clustering separates the cancer and normal samples. The significance of the misplaced samples will be discussed later.

Descriminant analysis allowed us to identify 20 peaks that could be used to predict disease classes in patients with cancer, benign disease, and normal prostate tissue with 90.5% accuracy. GeneSpring differentiated samples with benign, normal and cancer samples with 100% accuracy. It is worth noting that although the normal samples that clustered as cancer were mispredicted, they also showed to have elevated PSA levels that suggests early stage prostate cancer.

In conclusion, using GeneSpring for the analysis of protein spectra allowed us to meet our goal of identifying protein peaks that are markers for prostate disease in a matter of hours using a desktop computer.

This type of analysis could make serum biomarker testing a viable means to an early, accurate diagnosis of precancerous and prostate cancer cells, with the potential for early intervention and improved clinical outcomes. In addition, this type of analysis can be applied to other systems where serum biomarker classification can be used to supplement diagnosis.

References