Application of the GA/KNN method to SELDI proteomics data

Leping Li1,*, David M. Umbach1, Paul Terry2 and Jack A. Taylor2,3

1Biostatistics Branch, 2Epidemiology Branch and 3Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709, USA

Received on July 21, 2003; revised on December 15, 2003; accepted on December 16, 2003
Advance Access publication February 12, 2004

ABSTRACT
Summary: Proteomics technology has shown promise in identifying biomarkers for disease, toxicant exposure and stress. We show by example that the genetic algorithm/k-nearest neighbors method, developed for mining high-dimensional microarray gene expression data, is also capable of mining surface enhanced laser desorption/ionization–time-of-flight proteomics data.

Availability: The source code of the program and documentation on how to use it are freely available to non-commercial users at http://dir.niehs.nih.gov/dirbb/lifiles/softlic.htm
Contact: li3@niehs.nih.gov

INTRODUCTION
Recent advances in technologies for proteomics have enabled large-scale analysis of complex protein expression patterns, protein–protein interactions and post-translational modifications (Aebersold and Mann, 2003; Phizicky et al., 2003). One of these technologies, surface enhanced laser desorption/ionization–time-of-flight (SELDI–TOF) mass spectrometry, has shown great potential in identifying biomarkers of cancer from serum samples, thus promising early disease diagnosis and prevention [for a recent review, see Liotta et al. (2003)]. Recently, SELDI–TOF has been applied successfully to the identification of biomarkers for ovarian cancer (Petricoin et al., 2002a), breast cancer (Carter et al., 2002) and prostate cancer (Petricoin et al., 2002b).

Although significant progress has been made recently in technologies for proteomics, the development of tools for the analysis and interpretation of the resultant large amounts of data remains a challenge. Few methods have been reported in the literature for class discrimination based on SELDI mass spectra. One approach is via tree-based methods (Qu et al., 2002); another is a genetic algorithm (GA) approach whose details are proprietary (Petricoin et al., 2002a). Earlier, one of us developed a GA/k-nearest neighbors (GA/KNN) method for selecting a small number of discriminative genes for sample classification from the thousands of genes in a microarray expression analysis (Li et al., 2001a,b). Like high-dimensional expression data, SELDI–TOF mass spectroscopy data also consist of tens of thousands of mass/charge (m/z) ratios (analogous to genes) per specimen and an intensity level for each m/z ratio. Because of this similarity and the generality of the GA/KNN algorithm in pattern recognition, we believe that the GA/KNN algorithm can be useful in mining high-dimensional SELDI–TOF data. However, SELDI–TOF data do differ from gene expression data in important ways: both the number of specimens (hundreds per group compared with tens) and the number of variables are generally larger in SELDI–TOF data.

In this study, we demonstrate the applicability of the GA/KNN method to SELDI proteomics data analysis. We show by example that GA/KNN is capable of finding a few ions (m/z ratios) capable of reliable discrimination between cancer and unaffected serum specimens using a published SELDI dataset (Petricoin et al., 2002a).

METHODS
The details of the GA and KNN were described earlier (Li et al., 2001a,b) and can be found at http://dir.niehs.nih.gov/microarray/datamining. Briefly, the GA/KNN is a supervised multivariate classification method that selects many subsets of variables (e.g. m/z ratios for SELDI–TOF data or genes for microarray expression data) that discriminate between different classes of specimens. It employs a GA as a search tool to choose a relatively small subset of m/z ratios and uses KNN as a non-parametric pattern recognition method to evaluate the discriminative ability of the subset. For high-dimensional SELDI–TOF data with a paucity of samples, many distinct subsets of m/z ratios may be able to discriminate between different classes of specimens. Therefore, it is important to examine as many subsets of discriminative m/z ratios as possible. When a large number of such subsets has been obtained, the frequency with which m/z ratios
are selected into the subsets can be examined. The selection frequency should correlate with the relative predictive importance of \( m/z \) ratios for sample classification: the most frequently selected \( m/z \) ratios should be most discriminative, whereas the least frequently selected \( m/z \) ratios should be less informative. After ranking the \( m/z \) ratios according to frequency of selection, the top-ranked \( m/z \) ratios may be designated as the ‘best’ subset and used to classify unknown specimens in a validation set.

APPLICATION

To test its proteomics application, we applied the GA/KNN algorithm to a dataset (Petricoin et al., 2002a) that is publicly available at http://clinicalproteomics.steem.com/ppatterns.php. The data come from a SELDI–TOF analysis of the serum of 100 unaffected women and 100 patients who later developed ovarian cancer. Each specimen has intensity values obtained from SELDI mass spectra for 15 154 \( m/z \) ratios. Our goal, as in the original study, was to identify a few \( m/z \) ratios that may be used as predictive markers for ovarian cancer.

We applied the range transformation described by Petricoin et al. (2002a): for each \( m/z \) ratio, the transformed intensity was formed by taking the difference between the observed intensity and the minimum intensity and dividing that difference by the difference between the maximum and minimum intensities. The transformed intensities lie between 0 and 1 for each \( m/z \) ratio. These standardized data were subjected to GA/KNN analysis.

We randomly chose 50 from the 100 unaffected specimens and 50 from the 100 cancer specimens as the learning set and reserved the remainder as a validation set. The learning set was used to select 10 000 subsets of 20 \( m/z \) ratios each that jointly discriminate between unaffected and ovarian cancer specimens (using subsets of 10 or of 30 \( m/z \) ratios gave similar results). A subset of \( m/z \) ratios was considered discriminative when at least 90 of the 100 learning specimens were correctly classified (\( k = 5 \), consensus rule). The 15 154 \( m/z \) ratios were subsequently rank ordered according to the number of times each was selected into the 10 000 discriminative subsets as described above. Next, the top-ranked \( m/z \) ratios were used to classify \( k = 5 \), majority rule) each of the 100 specimens in the validation set as either an ovarian cancer or an unaffected specimen. This classification was repeated 100 times; first using the top \( m/z \) ratio then the top two ratios and so forth up to the top 100 \( m/z \) ratios as the ‘best’ subset. To assess the robustness of the result, we repeated the entire GA/KNN procedure 50 times. Each time, 50 cancer and 50 unaffected specimens were chosen randomly as the learning set and the remainder as the validation set.

Figure 1 shows results of the 50 runs. As the number of top-ranked \( m/z \) ratios increased, the percentage of correct prediction also increased and reached its plateau around 10.

This result suggests that 10 top-ranked \( m/z \) ratios are sufficient for an optimal discrimination between unaffected and ovarian cancer specimens for this dataset. Interestingly, the single most discriminative \( m/z \) ratio \( (m/z = 0.42) \) was able to predict correctly 83–94% of the specimens in repeated independent partitions into learning and validation sets. As expected, the 10 top-ranked \( m/z \) ratios from most of the 50 runs are identical \((0.42, 0.08, 0.07, 0.43, 0.05, 0.52, 444.08, 0.53, 0.09 \) and \( 0.25) \). On average, 97% \((93–100%) \) of the specimens in the validation set were correctly classified using 10 top-ranked \( m/z \) ratios; the corresponding rates were 95% \((range 90–100%) \) for the unaffected specimens and 98% \((range 90–100%) \) for the cancer specimens.

Surprisingly, all the top 10 most discriminative \( m/z \) ratios that we found were below 500. Because such low \( m/z \) ratios are regarded as likely reflecting the surface coatings and not serum proteins, investigators routinely omit low ratios from their analyses. Consequently, the specific \( m/z \) ratios that we found must be regarded skeptically as possibly reflecting non-biological experimental differences between the normal and cancer specimens. Sorace and Zhan (2003) drew attention to the same issue when analyzing a newer ovarian cancer dataset (dated July 8, 2002) from the same source as ours. To examine the classification result without the low \( m/z \) ratios, we removed all \( m/z \) ratios that were below 500 and repeated the GA/KNN analysis. On average, 90% \((78–96%) \) of the specimens were correctly classified using the 10 top-ranked \( m/z \) ratios \((1519.60, 1520.33, 1518.87, 1521.05, 1516.69, 1518.14, 1224.14, 533.10, 1343.83 \) and \( 3450.25) \).

To conclude, the GA/KNN method is capable of identifying discriminative markers from SELDI–TOF analysis of serum samples. As many proteomics initiatives in environmental toxicology and cancer biology begin, the GA/KNN method should be useful in discovering biomarkers of chemical exposure and disease.
ACKNOWLEDGEMENTS

We thank Pierre Bushel for his comments and Alex Merrick and Jennifer Madenspacher for insightful discussions.

REFERENCES


