Bioinformatics Approaches to Biomarker Discovery

BIOC 218/ BMI 231

FINAL PROJECT

March 2007

Seema Verma
seemav@stanford.edu

1. Introduction

The availability of the complete human genome has paved the way for the systematic understanding of human diseases. Recent technological advances in functional genomics and proteomics have fueled interest in identifying the biomarkers of complex diseases such as cancer and neurodegenerative diseases enabling a systems level analysis.

Functional genomics describes the use of large scale data produced by high throughput (HTP) technologies to understand the function of genes and other parts of the genome. With the help of high-throughput gene expression technologies, it is possible to analyze the expression of a large number of sequences in diseased and in normal tissues. The experimental approaches used to profile gene expression in complex human diseases include the representational differential display and microarrays together with real time Q-PCR for cross validation. The increasing size and complexity of the data generated by HTP methods provide challenges for researchers to extract the biologically relevant information. It is important that microarray technology and bioinformatics approaches be used in conjunction to facilitate biomarker discovery.

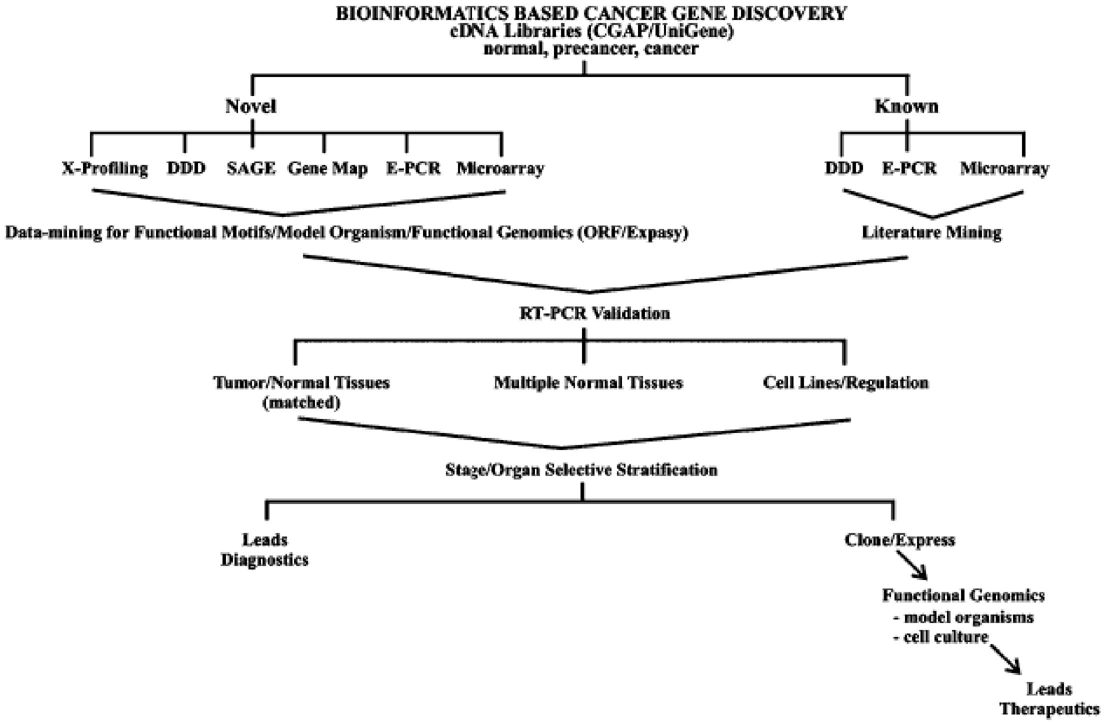

Fig. 1: A proposed approach to cancer gene discovery from the CGAP database is shown. Both novel and known ESTs are identified using multiple data-mining tools from this database. Further validation in the wet laboratory provides a rational for diagnostic and therapeutic target discovery. Adapted from Narayanan, R. (2007). "Bioinformatics approaches to cancer gene discovery." Methods Mol Biol **360**: 13-31.

An overall strategy for cancer gene discovery by using bioinformatics approaches is shown in Fig.1. Narayanan and co workers discovered two genes using data mining approaches with cancer genome (Narayanan 2007).

Genomics provided the blue print of possible gene products that are the main focus of proteomics. Proteomics would not be possible without genomics. While there are about 30,000 genes in the human genome (http://www.ncbi.nlm.nih.gov/genome/guide/human/), the protein complement of a cell or tissue, the proteome, is much larger and is also much more dynamic in nature. This is because most eukaryotic genes show alternative splicing of transcripts leading to different isoforms of a given protein. This, coupled with post-translational modification such as glycosylation, myristylation, and phosphorylation, leads to two or more effectively different proteins per gene.

Despite the availability of powerful genomics and transcriptomics technologies that are rapid discovery tools, one important shortcoming of these approaches is the lack of correlation between mRNA levels and changes in the protein expression. Many different technologies have been and are still being developed to collect the information contained in the proteins. Fig. 2 summarizes the current state of these technologies and their relationship to other discovery tools (Patterson and Aebersold 2003).

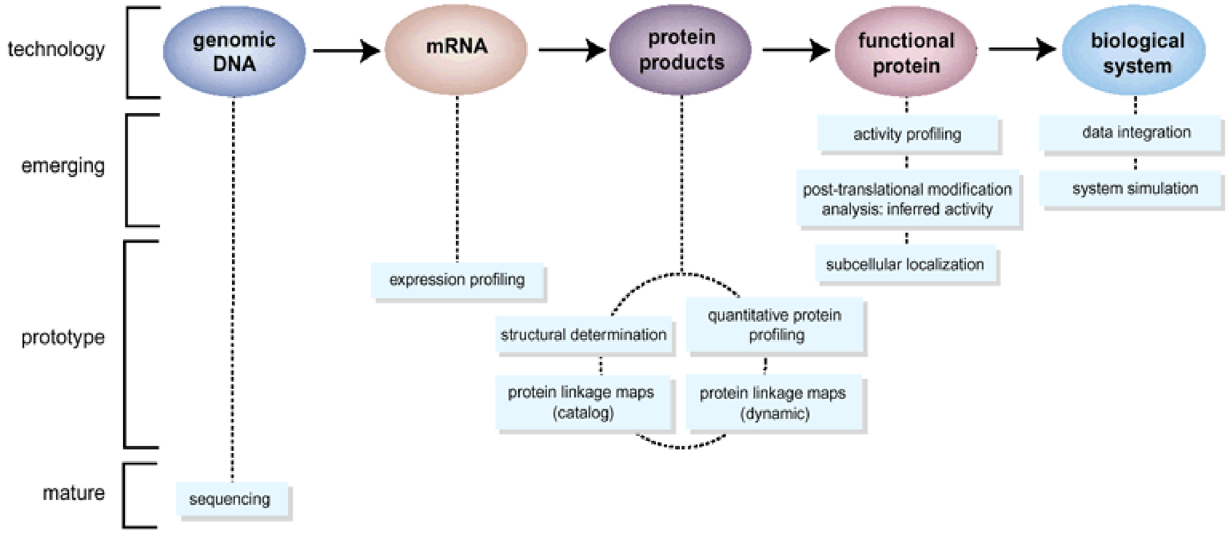

Fig. 2: Current status of proteomic technologies. Adapted from Patterson, S. D. and R. H. Aebersold (2003). "Proteomics: the first decade and beyond." Nat Genet **33 Suppl**: 311-23.

Most proteomics studies start with the fractionation of clinical samples from a case group and another set from a control group. Analysis of samples is carried out by mass spectrometry (MS) or 2-D gel. The data generated from these analyses is subjected to data mining approaches for complex pattern recognition resulting in the discovery of a set of biomarkers. A potential workflow of the proteomics process in biomarker discovery is shown in Fig.3.

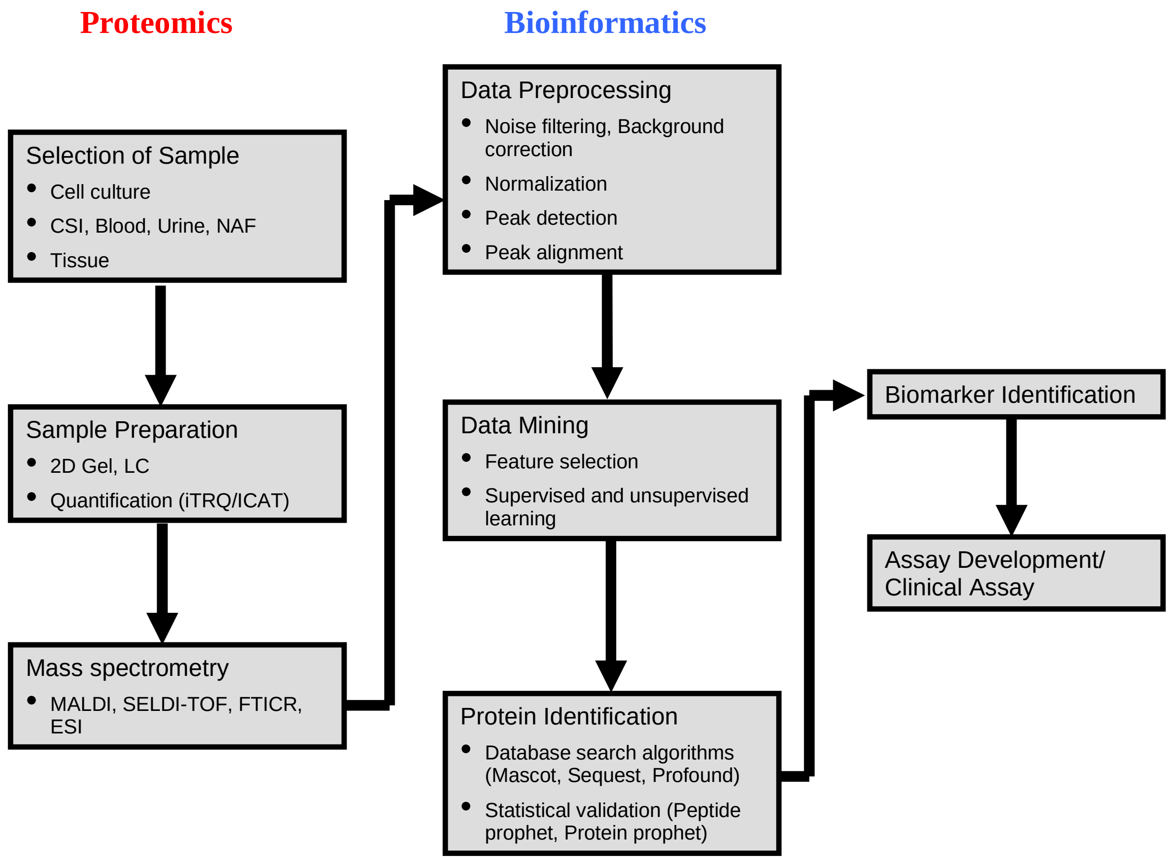
Proteomics
Bioinformatics
Selection of Sample
Cell culture
CSI, Blood, Urine, NAF
Tissue
Sample Preparation
2D Gel, LC
Quantification (iTRQ/ICAT)
Mass spectrometry
MALDI, SELDI-TOF, FTICR, ESI
Data Preprocessing
Noise filtering, Background correction
Normalization
Peak detection
Peak alignment
Data Mining
Feature selection
Supervised and unsupervised learning
Protein Identification
Database search algorithms (Mascot, Sequest, Profound)
Statistical validation (Peptide prophet, Protein prophet)
Biomarker Identification
Assay Development/ Clinical Assay

Fig. 3: Schematic representation of the proteomics process in biomarker discovery. Samples from groups are analyzed using various proteomics technologies. Sample fractionation is followed by analysis with mass spectrometry which generates millions of data points. These data points are subjected to preprocessing steps. The processed data is submitted to data mining approaches that result in a set of biomarkers that are identified by data search algorithms. Finally, the biomarkers are available for an HTP assay.

This review will critically analyze the recent developments in bioinformatics and data mining approaches in proteomics for biomarker discovery in complex diseases such as cancer and neurological disorder. Due to space constraints, I won't be covering candidate proteins in these diseases. In particular, this review will focus on important computational concepts and will outline the procedure for processing of MS data to obtain better quality results.

The main aim for investments in the development of proteomics is to develop advanced methods of disease diagnoses, understanding of the disease processes, and remedies and potential treatment of the disease at an early stage. Protein expression profiling is increasingly being used to discover, to validate, and to characterize biomarkers. Targeted and profiling approaches are being implemented to discover biomarkers. The former approach targets the selected set of disease related proteins. Profiling approach is an unbiased approach that does not rely on the prior information on the protein of interest. Recent advances in mass spectrometry and improved bioinformatics and statistical tools have revolutionized the biomarker discovery approach. In biomarker discovery, much of the efforts have been directed towards the development of strategies and platforms for quantitative protein profiling based upon the needs of different types of biological samples. The biomarker search can be performed on tissues, on body fluids, or on cultured cells. Body fluids may include urine, saliva, tears, sweat, and nipple aspirate fluid. The last may exhibit a lot of variation as compared to serum and cerebrospinal fluid (CSF). For most of the neurological disorders serum and CSF are used for proteomics or metabolomics analysis.

Techniques used in biomarker discovery include 2-D gel electrophoresis, gel free MS, and protein array technology. The more widely used approach, 2-D gel electrophoresis, which provides the capability to qualitatively and quantitatively resolve complex protein mixtures to unique spots, is a potential tool for biomarker discovery (Rai and Chan 2004). Main limitations of this method are its limited reproducibility and proteins that are expressed at low levels. These low levels may result in undetectable proteins which significantly limits the application of this method to clinical samples. The newly improved 2-D approach, differential in-gel electrophoresis (DIGE), has better reproducibility and throughput (Van den Bergh and Arckens 2005).

Gel-free mass spectrometry based approaches to biomarker discovery include Liquid Chromatography LC-MS (Wall, Kachman et al. 2001), Fourier-transform ion cyclotron resonance FTICR-MS or LC-FTICR AMT tag approach (Conrads, Anderson et al. 2000; Umar, Luider et al. 2007), surface-enhanced laser desorption/ionization SELDI-TOF-MS (Tang, Tornatore et al. 2004), and matrix-assisted laser desorption/ionization MALDI-TOF-MS (Reyzer and Caprioli 2005).

Profiling-based approaches are increasingly being used to discover and validate biomarkers using MS-based techniques. Most of the MS-based techniques for *peptide profiling* use two ionization techniques: MALDI and electrospray ionization (ESI). In MALDI, the mass of the anylate is estimated by time of the flight (TOF) analyzer. MALDI, when coupled with Fourier transform (FT-MS) enables high sensitivity and high mass accuracy measurements. In ESI mostly multiple protonated peptides are observed, whereas in MALDI only one protonated peptide is observed. Different analyzers can be used for ESI : e.g. TOF, quadrupole, ion trap, and FTMS. *Tryptic peptide profiling* is another method to discover biomarkers since enzymatic digestion significantly improves the resolution and the sensitivity of mass measurements. Both MALDI and ESI can be used as ionization techniques to study differentially expressed proteins that can be identified from complex protein mixtures. Dekker *et.al* demonstrated the use of MALADI-TOF-MS for tryptic profiling of CSF samples from 106 breast cancer patients and 45 control samples and found 164 differentially expressed peptides (Dekker, Boogerd et al. 2005). SELDI approach was used to identify ovarian cancer in serum which engendered enormous interest in profiling of proteins and peptides in body fluids (Petricoin, Ardekani et al. 2002). This approach was also applied to CSF for patients with neurodegenerative diseases (Lewczuk, Esselmann et al. 2003; Ruetschi, Zetterberg et al. 2005). CSF is a storehouse of naturally occurring peptide generated from neuropeptides; growth factors are also analyzed by more sensitive methods (Selle, Lamerz et al. 2005).

Data pre-processing

a. Background corrections

When exploring MS data, the spectral peaks produced from ionization of peptides and proteins are biologically relevant. A number of steps are involved in the data pre-processing to detect and to locate spectral peaks. These include spectrum calibration, base-line correction, smoothing, peak identification, intensity normalization, and peak alignment. The most important part of this process is the reduction and filtering of the raw data. Noise in spectra from chemical and electronic sources produce background signals, and it is important to check background correction before further analysis. Background fluctuations in MALDI and SELDI-TOF can create high background with low mass. Satten et al. proposed standardization and deionizing algorithms for background correction (Satten, Datta et al. 2004). Local smoothing methods have been utilized for baseline subtraction to remove high frequency noise, which may be apparent in MALDI-MS spectra (Wu, Abbott et al. 2003). Wu et al. (2003) used a local linear regression method to estimate the background intensity values, and then subtracted the fitted values from the local linear regression result.

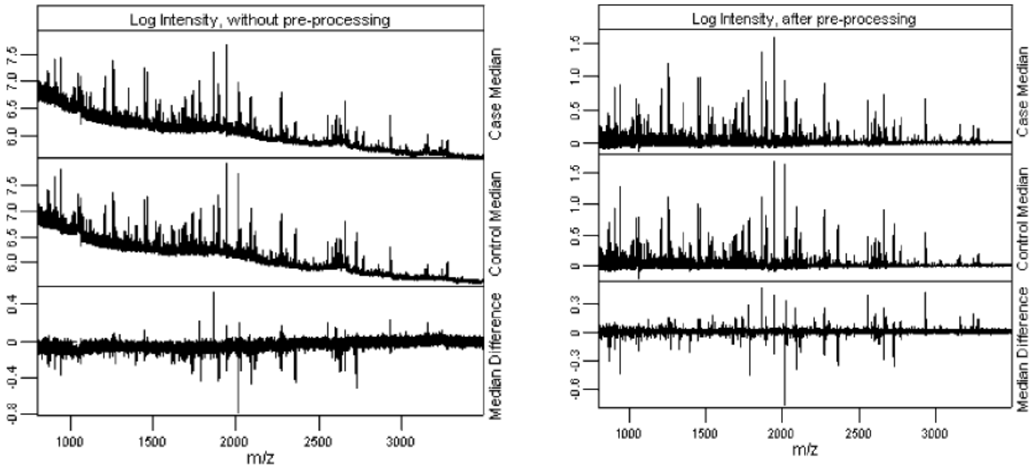

Fig. 4: Data preprocessing example. Left: original raw data. Right: mass spectrometry data after baseline correction. Adapted from Wu, B., T. Abbott, et al. (2003) Bioinformatics **19**(13): 1636-43.

Malyrenko et.al (2005) computed the response of SELDI-TOF data to increase in the charge and applied suitable digital filters for correction (Malyarenko, Cooke et al. 2005).

b. Peak detection

Once we have the background corrected spectra, the next step is peak detection. Peak detection helps to reduce the size of the spectral data and quantifies the set of peptides that are differentially expressed between different samples. For peak detection, peak finding algorithms are used. Coombes et al. (2003) proposed a method that initially uses the peak detection to obtain a preliminary list of peaks and then a baseline is calculated by excluding candidate peaks from the spectrum. The two steps are iterated and some peaks are further filtered depending on signal-to-noise ratio (Coombes, Fritsche et al. 2003). Another algorithm uses the entire peak to find the location on spectra by fitting Gaussian distribution to the peak (Kempka, Sjodahl et al. 2004). For low resolution spectra obtained from MALDI-TOF, the peak detection algorithm requires user input regarding the number of neighboring points and the intensity threshold value. False positive detections can be avoided by incorporating some additional constraints. High resolution algorithms like SNAP or THRASH proposed the picking of one isotopic variant of a peptide peak and treating it as distinct from background signals by analyzing the isotopic distribution (Horn, Zubarev et al. 2000).

c. Peak alignment

After detecting the peak, the spectrum is aligned. Alignment algorithms typically involve either: (i) maximizing some objective function over a parametric set of transformations, or (ii) nonparametric alignment, by way of dynamic programming (DP), or (iii) some combination of these methods. Zhang et al. (2005) designed a two-step alignment algorithm recognizing peaks generated by the same peptide but detected in different samples and addresses systematic retention time shift. In gross alignment, all possible significant peaks were first identified. A significant peak refers to a peak that is present in every sample and is most intense in certain m/z and retention-time range (Zhang, Asara et al. 2005). After gross alignment, microalignment identifies peaks of the same molecule in different datasets. So, gross alignment adjusts the overall retention time drift between samples, while the microalignment focuses on the local complexity and aligns peaks together. Randolph and Yasui (2004) used coarse scale-specific peaks, extracted by multiscale wavelet decomposition, to align MALDI data along the m/z axis. They used a coarse-to-fine method to first align peaks at a dominant scale and then refine the alignment of other peaks at a finer scale. But it is unclear if the multiscale approach is biologically reasonable (Randolph and Yasui 2006). Dynamic programming (DP) based approaches have also been proposed (Nielsen 1998). Unlike microarray, in MS data analysis, one-to-one correspondence between two data sets does not always exist. It also remains unclear how DP can identify and ignore outliers during the matching. Eilers (2004) proposed a parametric model for the warping function when aligning chromatograms. The parameters of the warping function are easily interpolated, allowing alignment of batches of chromatograms based on warping functions for a limited number of calibration samples a parametric warping model with polynomial functions or spline functions to align chromatograms (Eilers 2004).

Data normalization

The normalization step helps to reduce variation due to experimental noise from systemic effect between samples, e.g., from varying amounts of applied protein, degradation over time in the sample, or change in the column or sensitivity of instrument. Normalization of MS data can be performed either by coercing the *m/z* intensity values to be comparable across experiments (low-level processing), or by altering the peak abundance to be comparable (mid-level processing). In general, one aims to normalize not only replicates, but also experimental data of distinct biological origin, such as serum profiles from cancer patients and healthy case controls. The underlying assumption behind normalization is that the overall MS abundance of all features (peaks or time-*m/z* pairs), or subset(s) of these, should be equal across different experiments (Listgarten and Emili 2005). Global normalization refers to cases where all features are simultaneously used to determine a single normalization factor between two experiments, while local normalization refers to cases where a subset of features are used at a time (different subsets for different parts of the data). Wang and co-workers (2006) proposed a two-step normalization procedure. A global normalization is followed by a probability model to investigate the intensity-dependent missing events and provides possible substitutions for the missing values (Wang, Tang et al. 2006). Baggerly et.al (2004) used normalized intensities NV_i produced by the SELDI method. For a single spectrum, V_i denotes the raw intensity at the i-th m/z value. $V_{\min}$ and $V_{\max}$ denote the smallest and the largest observed intensities in the spectrum, respectively (Baggerly, Morris et al. 2004). Then the normalized intensity NV_i is given by

$$NV_i = \frac{V_i - V_{\min}}{V_{\max} - V_{\min}}$$

Data mining

Datapoints obtained from the data preprocessing step represent potential biomarkers. Many profiling studies aim to find proteomic patterns that can discriminate between different biological conditions. In order to properly assign statistical significance to candidate biomarkers, or any changes in apparent protein abundance, it is important to understand the patterns of variability. Before subjecting the data to data mining algorithms, a feature selection step is used which can be performed on raw data or the detected peaks using unsupervised learning approaches (approaches do not take into account class labels; analogous to clustering) or supervised learning approaches (approaches accounts for class labels; analogous to classification) which are discussed below. Yu et.al implemented a random forest algorithm to find markers that can best discriminate cases from control sample (Wu, Abbott et al. 2003; Yu, Wu et al. 2006). Pratapa et.al compared feature selection with Fisher discriminant ratio (FDR), followed by classification accuracy of a linear SVM versus joint feature selection and classification with Bayesian sparse multinomial logistic regression (SMLR). The SMLR approach outperformed FDR and SVM, but both were effective in achieving good diagnostic accuracy with a small number of features (Pratapa, Patz et al. 2006). Once the features are selected, the data undergoes transformation due to high variance in a given input variable. Methods include log transformation, square root transformation, or linear and logarithmic scaling (Stein 1999; Anderle, Roy et al. 2004). Classification trees ignore variance and therefore transformation is not essential.

Many profiling studies aims to find proteomic patterns that can discriminate between different biological conditions. Data mining is an important element in databases that can be used to extract the hidden information by supervised or unsupervised learning methods.

Unsupervised learning methods

Unsupervised approaches are simplest routine approach to visualize the distribution of data. These approaches include k-means clustering, principal component analysis (PCA), and hierarchical clustering which can be used a basis for feature selection. PCA maps high dimensional data by creating eigenvalues. Each linear combination or principal component is a weighted sum of the amplitude at each m/z value. The feature selections of PCA are present in the top principal components which separates the samples into homogeneous clusters and can be visualized in 2D or in 3D plots in which the calculated values for top principal components serve as x, y, and z axes (Duda RO 2000). The PCA approach was used to rank peak intensities within each spectrum and applied on cervical (Hellman, Alaiya et al. 2004) and borderline ovarian cancers (Alaiya, Franzen et al. 2002).

Hierarchical clustering (HC) is another powerful data mining method for initial exploration of proteomic data. HC begins by assigning each sample to its own cluster. It further calculates similarity scores or distance matrices between sample and places samples that are close to each other. HC algorithms may differ in calculating distance matrices. Two way clustering algorithm was used to differentiate cancerous from non-cancerous cells and human CSF (Poon, Yip et al. 2003; Hu, Malone et al. 2005; Meunier, Dumas et al. 2007).

Supervised Learning Approaches

Supervised learning techniques require class labels such that training can occur on data obtained from a subset of the provided samples. The two types of variables in this exercise are predictor variables (intensity at m/z values or peak intensity) and a response variable (disease). The straight approach to identify the differenced between two group would be a t-test using a supervised method. Unlike Welch t-test, Mann-Whitney test assumes equal variance between the two groups. The t-test has some limitations like by calculating t-statistics for each peak, Multiple testing can give more number of variables, and these calculations assume that measurements are independent of each other. Bonferroni correction reduces the impact of multiple testing procedure (Belknap 1992).

Classification algorithms can be used for feature selection and classification. Such algorithms include genetic algorithms, decision trees, and neural networks. The aim of genetic algorithms (GA) is to extract a model by creating chromosomes of input variables (m/z values) and iteratively recombining chromosomes and mutating genes. More specifically, this mathematical model relates the protein abundance with the presence of a certain gene. This process progressively becomes more difficult as the number of variables grow. Input variables that satisfy fitness function are kept and the rest are discarded through computational evolution. GA have been used in various MS datasets (Jeffries 2004; Shadforth, Crowther et al. 2005). Petricoin et.al used GA and self organizing maps (SOM) and applied these to the development of diagnostic patterns for ovarian and prostrate cancers to find a good set of predictive SELDI *m/z* values (Petricoin, Ornstein et al. 2002)**.**

Decision trees start with the entire sample dataset and create a decision rule that divides the entire sample dataset into two homogeneous groups. The decision rule examines the input variable and partitions the dataset into branches based on the less than function. New branch is then examined for homogeneity and can be subdivided employing a new rule. Each final node is labeled as a class using majority rule based on the training samples. Adam *et.al* used this classification technique to an MS dataset of prostrate cancer patients resulting in 96% accuracy, 83% sensitivity, and 97% specificity. Decision trees produce interpretable and applicable decision rules for classifying samples and have been used extensively in the analysis of proteomic data (Geurts, Fillet et al. 2005; Yang, Xiao et al. 2005; Albitar, Potts et al. 2006)

Artifical neural networks (ANN) are based on the way the human brain processes information. Neurons integerate information obtained from different inputs which could be outside world (primary level data) or previously integrated data (other neurons). Most neural networks feed forward, i.e., information flow is unidirectional, starting with an input layer flowing through n layer of neurons and finally to the output. Training the neural network involves decreasing the error rate by adjusting model parameters, i.e., assigning weights to input function, activation threshold of each neuron, and computation function performed by each neuron. Due to their low error rates, artificial neural network algorithms have been applied to analyze mass spectra for cancer and neurodegenerative diseases (Ball, Mian *et al*. 2002; Poon, Yip et al. 2003; Di Luca, Grossi et al. 2005; Gobel, Vorderwulbecke *et al*. 2006; Ru, Zhu et al. 2006).

Support Vector machines (SVMs) is another machine learning approach which is applied to MS data analysis. SVMs operate first by distributing the sample in n-dimensional space and then by finding a hyper space that attempts to split the cases from controls samples (Burges 1998). Numerous studies have used SVMs for MS data analysis (Li, Zhang et al. 2002; Zhang, Bast et al. 2004; Yu, Zheng et al. 2005). Li *et al.* selected 10 *m/z* values as features in three SELDI datasets trying both a t-test filter and a genetic algorithm. These were used in conjunction with an SVM classifier, where the choice of kernel was reported to have little effect (Li, Zhang et al. 2002). Wagner *et al.* used each of: *k*-nearest-neighbor (*k* = 6, Mahalanobis distance), support vector machine (SVM) with linear kernel (LDA), and quadratic discriminant analysis (QDA) to classify MALDI data selecting the top 3\N15 peaks as features with an *F* statistic (Wagner, Naik et al. 2003).

Comparative classifiers: Wu *et al.* compared the performance of LDA, QDA, *k*-nearest neighbor (*k* = 1\N3, Euclidean metric), bagging and boosting classification trees, SVM (kernel not specified), and RF on MALDI data, using both a t-test rank and the by-product of the RF algorithm for *m/z* feature selection (15 and 25 features). Overall, no substantive differences in performance were reported, with QDA marginally best, although different error estimators (cross-validation or bootstrap) were used for different classifiers, complicating the interpretation (Wu, Abbott et al. 2003).

Many investigators have analyzed MS data for cancer and neurological disorders using the ProPeak and Ciphergen Peaks 2.1 software combined with visual analysis (Zhu, Wang et al. 2003; Ranganathan, Williams et al. 2005; Ranganathan, Williams et al. 2005; Guerreiro, Gomez-Mancilla et al. 2006; Huang, Leweke et al. 2006; Lakhan 2006). The software detects biomarker peaks or features that differentiate spectra of cancer and non-canerous patients.

Conclusions:

Bioinformatics approaches are critical for effectively mining high-dimensional data to provide insights into disease biology. Data preprocessing such as background correction and spectrum alignment are critical issues before data mining. High dimensional data needs to be reduced to fewer variables using feature selection. Many algorithms exist to mine large datasets, but no specific approach is ideal or applicable to all study designs. For data mining, best approach would be to utilize feature selection algorithm with cross validation. It is better to utilize different approaches in parallel to arrive at a final algorithm. With increasing availability of public data, rigorous comparisons of data preprocessing and data mining approaches are needed. Most of the proteomics studies are performed on small populations. It is possible that small sample size may result in potential biomarkers failing the validation test. MS is increasingly being used to analyze complex protein mixtures to recognize biomarker patterns. SELDI based profiling appears to successfully detect some previously unknown proteins. Also, there is evidence that biomarker patterns can be found that can differentiate cancerous and normal individuals. Finally, it is anticipated that existing and emerging computational data mining approaches along with rigorous and systematic evaluation, will help to unleash the full biological potential of proteomic profiling.

References:

Alaiya, A. A., B. Franzen, et al. (2002). "Molecular classification of borderline ovarian tumors using hierarchical cluster analysis of protein expression profiles." Int J Cancer **98**(6): 895-9.

Albitar, M., S. J. Potts, et al. (2006). "Proteomic-based prediction of clinical behavior in adult acute lymphoblastic leukemia." Cancer **106**(7): 1587-94.

Anderle, M., S. Roy, et al. (2004). "Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum." Bioinformatics **20**(18): 3575-3582.

Baggerly, K. A., J. S. Morris, et al. (2004). "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments." Bioinformatics **20**(5): 777-85.

Ball, G., S. Mian, et al. (2002). "An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers." Bioinformatics **18**(3): 395-404.

Belknap, J. K. (1992). "Empirical estimates of Bonferroni corrections for use in chromosome mapping studies with the BXD recombinant inbred strains." Behav Genet **22**(6): 677-84.

Burges, C. (1998). "A tutorial on support vector machines for pattern recognition." Data mining knowledge discovery **2**: 121-167.

Conrads, T. P., G. A. Anderson, et al. (2000). "Utility of accurate mass tags for proteome-wide protein identification." Anal Chem **72**(14): 3349-54.

Coombes, K. R., H. A. Fritsche, Jr., et al. (2003). "Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization." Clin Chem **49**(10): 1615-23.

Dekker, L. J., W. Boogerd, et al. (2005). "MALDI-TOF mass spectrometry analysis of cerebrospinal fluid tryptic peptide profiles to diagnose leptomeningeal metastases in patients with breast cancer." Mol Cell Proteomics **4**(9): 1341-9.

Di Luca, M., E. Grossi, et al. (2005). "Artificial neural networks allow the use of simultaneous measurements of Alzheimer disease markers for early detection of the disease." J Transl Med **3**: 30.

Duda RO, H. P., Stork,DG (2000). Pattern classification. New York, USA, Wiley.

Eilers, P. H. (2004). "Parametric time warping." Anal Chem **76**(2): 404-11.

Geurts, P., M. Fillet, et al. (2005). "Proteomic mass spectra classification using decision tree based ensemble methods." Bioinformatics **21**(14): 3138-45.

Gobel, T., S. Vorderwulbecke, et al. (2006). "New multi protein patterns differentiate liver fibrosis stages and hepatocellular carcinoma in chronic hepatitis C serum samples." World J Gastroenterol **12**(47): 7604-12.

Guerreiro, N., B. Gomez-Mancilla, et al. (2006). "Optimization and evaluation of surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry for protein profiling of cerebrospinal fluid

Proteomic profiling of cerebrospinal fluid identifies biomarkers for amyotrophic lateral sclerosis." Proteome Sci **4**(5): 7.

Hellman, K., A. A. Alaiya, et al. (2004). "Protein expression patterns in primary carcinoma of the vagina." Br J Cancer **91**(2): 319-26.

Horn, D. M., R. A. Zubarev, et al. (2000). "Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules." J Am Soc Mass Spectrom **11**(4): 320-32.

Hu, Y., J. P. Malone, et al. (2005). "Comparative proteomic analysis of intra- and interindividual variation in human cerebrospinal fluid." Mol Cell Proteomics **4**(12): 2000-9.

Huang, J. T., F. M. Leweke, et al. (2006). "Disease biomarkers in cerebrospinal fluid of patients with first-onset psychosis." PLoS Med **3**(11): e428.

Jeffries, N. O. (2004). "Performance of a genetic algorithm for mass spectrometry proteomics." BMC Bioinformatics **5**: 180.

Kempka, M., J. Sjodahl, et al. (2004). "Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry." Rapid Commun Mass Spectrom **18**(11): 1208-12.

Lakhan, S. E. (2006). "Schizophrenia proteomics: biomarkers on the path to laboratory medicine?" Diagn Pathol **1**: 11.

Lewczuk, P., H. Esselmann, et al. (2003). "The amyloid-beta (Abeta) peptide pattern in cerebrospinal fluid in Alzheimer's disease: evidence of a novel carboxyterminally elongated Abeta peptide." Rapid Commun Mass Spectrom **17**(12): 1291-6.

Li, J., Z. Zhang, et al. (2002). "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer." Clin Chem **48**(8): 1296-304.

Listgarten, J. and A. Emili (2005). "Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry." Mol Cell Proteomics **4**(4): 419-34.

Malyarenko, D. I., W. E. Cooke, et al. (2005). "Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques." Clin Chem **51**(1): 65-74.

Meunier, B., E. Dumas, et al. (2007). "Assessment of hierarchical clustering methodologies for proteomic data mining j. Proteome res. 2007, 6 (1), 358-366." J Proteome Res **6**(3): 1215.

Narayanan, R. (2007). "Bioinformatics approaches to cancer gene discovery." Methods Mol Biol **360**: 13-31.

Nielsen, N.-P. V., Carstensen, J. M., and Smedsgaard, J. (1998). "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping." J. Chromatogr. A **805**: 17 –35.

Patterson, S. D. and R. H. Aebersold (2003). "Proteomics: the first decade and beyond." Nat Genet **33 Suppl**: 311-23.

Petricoin, E. F., 3rd, D. K. Ornstein, et al. (2002). "Serum proteomic patterns for detection of prostate cancer." J Natl Cancer Inst **94**(20): 1576-8.

Petricoin, E. F., A. M. Ardekani, et al. (2002). "Use of proteomic patterns in serum to identify ovarian cancer." Lancet **359**(9306): 572-7.

Poon, T. C., T. T. Yip, et al. (2003). "Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes." Clin Chem **49**(5): 752-60.

Pratapa, P. N., E. F. Patz, Jr., et al. (2006). "Finding diagnostic biomarkers in proteomic spectra." Pac Symp Biocomput: 279-90.

Rai, A. J. and D. W. Chan (2004). "Cancer proteomics: Serum diagnostics for tumor marker discovery." Ann N Y Acad Sci **1022**: 286-94.

Randolph, T. W. and Y. Yasui (2006). "Multiscale processing of mass spectrometry data." Biometrics **62**(2): 589-97.

Ranganathan, S., E. Williams, et al. (2005). "Proteomic profiling of cerebrospinal fluid identifies biomarkers for amyotrophic lateral sclerosis." J Neurochem **95**(5): 1461-71.

Ranganathan, S., E. Williams, et al. (2005). "Proteomic profiling of cerebrospinal fluid identifies biomarkers for amyotrophic lateral sclerosis

Application of serum SELDI proteomic patterns in diagnosis of lung cancer." J Neurochem **95**(5): 1461-71.

Reyzer, M. L. and R. M. Caprioli (2005). "MALDI mass spectrometry for direct tissue analysis: a new tool for biomarker discovery." J Proteome Res **4**(4): 1138-42.

Ru, Q. C., L. A. Zhu, et al. (2006). "Label-free semiquantitative peptide feature profiling of human breast cancer and breast disease sera via two-dimensional liquid chromatography-mass spectrometry." Mol Cell Proteomics **5**(6): 1095-104.

Ruetschi, U., H. Zetterberg, et al. (2005). "Identification of CSF biomarkers for frontotemporal dementia using SELDI-TOF." Exp Neurol **196**(2): 273-81.

Satten, G. A., S. Datta, et al. (2004). "Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens." Bioinformatics **20**(17): 3128-36.

Selle, H., J. Lamerz, et al. (2005). "Identification of novel biomarker candidates by differential peptidomics analysis of cerebrospinal fluid in Alzheimer's disease." Comb Chem High Throughput Screen **8**(8): 801-6.

Shadforth, I., D. Crowther, et al. (2005). "Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines." Proteomics **5**(16): 4082-95.

Stein, S. E. (1999). "An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data." Journal of the American Society for Mass Spectrometry **10**(8): 770-781.

Tang, N., P. Tornatore, et al. (2004). "Current developments in SELDI affinity technology." Mass Spectrom Rev **23**(1): 34-44.

Umar, A., T. M. Luider, et al. (2007). "NanoLC-FT-ICR MS improves proteome coverage attainable for approximately 3000 laser-microdissected breast carcinoma cells." Proteomics **7**(2): 323-9.

Van den Bergh, G. and L. Arckens (2005). "Recent advances in 2D electrophoresis: an array of possibilities." Expert Rev Proteomics **2**(2): 243-52.

Wagner, M., D. Naik, et al. (2003). "Protocols for disease classification from mass spectrometry data." Proteomics **3**(9): 1692-8.

Wall, D. B., M. T. Kachman, et al. (2001). "Isoelectric focusing nonporous silica reversed-phase high-performance liquid chromatography/electrospray ionization time-of-flight mass spectrometry: a three-dimensional liquid-phase protein separation method as applied to the human erythroleukemia cell-line." Rapid Commun Mass Spectrom **15**(18): 1649-61.

Wang, P., H. Tang, et al. (2006). "Normalization regarding non-random missing values in high-throughput mass spectrometry data." Pac Symp Biocomput: 315-26.

Wu, B., T. Abbott, et al. (2003). "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data." Bioinformatics **19**(13): 1636-43.

Yang, S. Y., X. Y. Xiao, et al. (2005). "Application of serum SELDI proteomic patterns in diagnosis of lung cancer." BMC Cancer **5**: 83.

Yu, J. K., S. Zheng, et al. (2005). "An integrated approach utilizing proteomics and bioinformatics to detect ovarian cancer." J Zhejiang Univ Sci B **6**(4): 227-31.

Yu, W., B. Wu, et al. (2006). "MALDI-MS data analysis for disease biomarker discovery." Methods Mol Biol **328**: 199-216.

Zhang, X., J. M. Asara, et al. (2005). "Data pre-processing in liquid chromatography-mass spectrometry-based proteomics." Bioinformatics **21**(21): 4054-9.

Zhang, Z., R. C. Bast, Jr., et al. (2004). "Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer." Cancer Res **64**(16): 5882-90.

Zhu, W., X. Wang, et al. (2003). "Detection of cancer-specific markers amid massive mass spectral data." Proc Natl Acad Sci U S A **100**(25): 14666-71.